THE INTERNATIONAL JOURNAL OF INFORMATION, DIVERSITY, & INCLUSION

Bridging LGBT+ Content Gaps Across Wikipedia Language Editions

Marc Miquel Ribé, Wikimedia Foundation, USA
Andreas Kaltenbrunner, ISI Foundation, Italy
Jeffrey M. Keefer, New York University, USA

Abstract

In the past several years, the Wikimedia Movement has become more aware of the lack of representation of specific communities, that is, content gaps. Next to geographical and gender-related initiatives, the LGBT+ Wikimedia community has organized to create LGBT+ content encompassing (among other topics) biographies, events, and culture. In this paper, we present a computational approach to collecting and analyzing LGBT+ articles. We selected 14 Wikipedia language editions to study the coverage of LGBT+ content in general, its visibility in the list of Featured Articles, and its overlap with the local content of the Wikipedia language editions. Results show that a considerable part of potentially LGBT+ related content exists across Wikipedia language editions; however, this relation is not evident in each language edition. In this sense, closing the LGBT+ content gap is about creating articles and making connection to the topic visible in already existing articles. We also analyze the frequency of biographies of persons with non-heterosexual sexual orientations. We find that even though they represent only a small share of all biographies, they are a bit more frequent among the Featured Articles. When taking into account all the LGBT+ biographies of the different languages, English context celebrities are the most visible. While part of the LGBT+ content is related to each language edition's local context, it tends to be less contextualized than the entire language editions. This indicates the possibility of growing LGBT+ content in each Wikipedia language edition by representing its most immediate LGBT+ local context. We propose a dashboard tool to find relevant LGBT+ articles across language editions and start bridging the gaps. Finally, we conclude this study by presenting recommendations for the next steps amongst the Wikipedia communities to fill some of these gaps.

Keywords: content diversity; LGBT+; online communities; Wikipedia

Publication Type: research article

Introduction

LGBT+ Information Online

Over the past decades, there has been a growth of LGBT+ (Lesbian, Gay, Bisexual, Transexual, and other sexual identities) presence online. Social networks, and more generally online spaces, have become opportunities to self-express LGBT+ identities (Cooper & Dzara, 2010; Pullen & Cooper, 2010; Blackwell et al., 2016), as well as valuable tools to promote LGBT+ agendas by circumventing cultural and social barriers and overcoming

ISSN 2574-3430, https://jps.library.utoronto.ca/index.php/ijidi
DOI: 10.33137/ijidi.v5i4.37270

geographic distances (Ayoub & Brzezińska, 2016; Soriano, 2014). In addition, the appearance of online spaces has been useful to the LGBT+ community to support their activism and the generation, archiving, and access to information of interest (Cocciolo, 2017).

The online space Wikipedia has compiled one of the largest knowledge repositories on the Internet. By giving free access to "the sum of human knowledge", volunteers create articles about any topic. While libraries present social opportunities locally (Mehra & Srinivasan, 2007), Wikipedia has become an essential information resource open to everyone and available in more than 300 language editions. It is used in fact-checking, education, and news sources, among many other contexts (Okoli et al., 2014).

Rather than a substitute for libraries, Wikipedia is a general source of information that traditional knowledge-based institutions can nurture (Doyle, 2018; Phetteplace; 2015). As several authors point out Wikipedia can also be used to enhance the visibility of digitized archival assets (Szajewski, 2013; Galloway & DellaCorte, 2014; Cooban, 2017). Librarians and information professionals have an essential role in creating and expanding the articles and increasing the number of citations.

Campaigns like the GLAM (Galleries, Libraries, Archives, and Museums') (Wikipedia contributors, 2021a) initiative help cultural institutions in sharing their collections with the world through collaborative projects with seasoned Wikipedia editors. For the case of LGBT+ information, Wexelbaum (2019) examined the under-representation of librarians in global LGBT+ Wikipedia engagement efforts and Wikipedia initiatives in general, as well as the barriers that librarians face in becoming active Wikipedians (i.e., the volunteers who contribute to Wikipedia by editing its pages) (Wikipedia contributors, 2021b).

Wikipedia has gained much attention in medical studies (Herbert et al., 2015). Traditional high-impact medical and multidisciplinary journals are extensively cited in Wikipedia medical articles, indicating that the articles have robust underpinnings (Jemielniak et al., 2019). In addition, Wikipedia health-related articles can support decision-making by LGBT+ youth, given that some youth-oriented LGBT+ online communities and websites may provide inaccurate health information about the risks of infection and time for HIV seroconversion (Hawkins & Watson, 2017).

Wikipedia is often the first source readers access to get information, and in some contexts, it is the only source consulted for LGBT+ topics (Wexelbaum et al., 2015). For this reason, engaging information professionals who have expertise in topics related to LGBT+ and helping them become Wikipedians should be a priority. But how do the communities of the different Wikipedia language editions organize around creating LGBT+ information? How do they engage librarians and archivists, among others, to contribute to it?

Wikimedia LGBT+

To foster participation and be more effective at creating content, Wikipedians organize themselves around online and offline spaces with different levels of formality that range from time-bound events like edit-a-thons and contests to spaces like Wikiprojects and organizations (affiliates). For example, the first organized space to create LGBT-related articles in Wikipedia was the Wikiproject "LGBT studies" (Wikipedia contributors, 2021c) in the English Wikipedia, created in 2006 to identify, categorize, and create new LGBT queer articles on Wikipedia.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Wikiprojects are spaces of coordination within Wikipedia where editors can list articles to be created. Wikipedians categorize entries, review them, and identify potential new "stubs" that would require more work. For example, the "LGBT studies" Wikiproject exists in 28 Wikipedia language editions. In English, the showcase of articles created through the coordination of the project includes 400 "Good articles" and 100 "Featured articles," which have been through a review process to receive such distinctions (see Appendix A for more explanations on some of the Wikipedia specific terms used in this article).

Differently from a Wikiproject, an edit-a-thon is an activity that gathers several Wikipedians together in a physical or online place intending to create articles on a specific topic within a set period of time. These often occur in relation to cultural institutions or around set themes. For example, edit-a-thons have been organized in libraries, archives, and museums to leverage their collections and digitize LGBT+ cultural heritage to Wikipedia (Wexelbaum et al., 2015).

Edit-a-thons to improve the LGBT+ content in several Wikipedia language editions are usually organized by Wikimedia Movement affiliates. These are typically independent non-profit organizations that can use Wikipedia trademarks publicly and receive funding to organize events. The Wikimedia Movement is the totality of people, activities, and values which revolve around Wikimedia projects like Wikipedia ("Wikimedia movement," 2021). Since 2014, an affiliate named "Wikimedia LGBT+" has aimed to support the LGBT+ community and represent LGBT+ content across Wikimedia projects. Its mission is to "create and expand the content of interest to LGBT+ communities on Wikimedia projects, and to increase the overall quality of such content in all languages."

While Wikimedia LGBT+ is the only affiliate that focuses on LGBT+ content, and its working language is English, there are ten affiliates to bridge the gender gap in 4 different languages (French, English, Spanish, and Italian). Affiliates like "Whose Knowledge?" support intersectional participation and content in Wikipedia engagement and promote those who identify as women, LGBT+, people from the global south, and all those interested in addressing systemic bias on Wikimedia projects.

Affiliates like Wikimedia LGBT+ are a stable infrastructure to support the continuity of certain events and contests over time. For example, on a global scale, "Wiki Loves Pride" is a yearly campaign that started in 2014 to expand and improve LGBT+ content across several Wikimedia projects and organizes meetups and edit-a-thons in many countries. Most activities of Wiki Loves Pride take place between June and October, traditionally the months when lesbian, gay, bisexual, and transgender communities worldwide celebrate LGBT+ culture and history.

In 2021, Wikimedia LGBT+ has also organized the online conference "Queering Wikipedia 2021 User Group Working Days" to discuss internal operations and their participation in the Wikimedia Movement strategy conversations. The importance of these conversations is critical, as it is the venue where Wikipedians decide on the priorities for the Movement with the 2030 horizon. One of the strategic goals defined in these conversations is "knowledge equity" (Strategy/Wikimedia movement/2017/Direction, 2021), which calls to "counteract structural inequalities to ensure a just representation of knowledge and people in the Wikimedia movement."

Concerning this goal, the Wikimedia Foundation, the main organization in the Wikimedia Movement, has approved a Universal Code of Conduct, aiming to guarantee the safety of Wikipedians. Safety is an important aspect, considering LGBT+ Wikipedia volunteers may not feel

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

safe in their communities, especially in countries where LGBT+ content is taboo or banned, and it is not allowed or difficult to express an LGBT+ identity openly (Wexelbaum et al., 2015). Creating such code and past board resolutions against homophobia shows that safety is crucial for the Wikimedia Foundation. Furthermore, it communicates the importance of Wikipedia spaces becoming more welcoming to the LGBT+ community.

Measuring the LGBT+ Content Gaps

LGBT+ content is the product of many online and offline activities organized by Wikiprojects, Wikimedia affiliates, and the Wikimedia Foundation. However, to date, there is no precise counting of the available LGBT+ content in each Wikipedia language edition. Instead, editors choose new topics based on what they observe in Wikipedia and without knowing whether there are enough articles on a topic or not. Therefore, we can say that detecting potential content gaps related to LGBT+ after browsing the different categories occurs manually.

In contrast, for the Gender Gap, there exist tools that allow a user to quantify the number of women in relation to the total number of biographies in a Wikipedia language edition, accumulated or created in a specific period (Konieczny & Klein, 2018). While the proportion of women is far from parity (usually ranges from 15% to 20%), having a number encourages the different affiliates and groups of editors that prioritize closing this gap to keep working. Similarly, other content gaps like the Culture Gap and the Geography Gap have also been measured and monitored in dashboards (Miquel-Ribé & Laniado, 2020; Redi et al., 2020).

The creation of metrics to measure the LGBT+ gap has been claimed as a priority to the Wikimedia LGBT+ Affiliate (Wikimedia LGBT+/Portal, 2021). According to its page, one of the two ways in which it will fulfill its mission is to "create, collect, process, and present the sorts of metrics which describe usage statistics and quality of the content of LGBT+ interest on Wikimedia projects." However, differently from other content gaps such as the Gender or Geography gaps, LGBT+ Wikipedians need to consider one extra dimension: not only it is important to have articles dedicated to topics of interest to or about LGBT+ people, but also that they are presented publicly as related to LGBT+.

By taking a look at the LGBT+ Portal (Wikipedia contributors, 2021d) and at the Wikiproject "LGBT studies," we rapidly see that the scope of topics that can relate to LGBT+ is wide: history of the LGBT+ rights, social attitudes, LGBT+ culture including movies, literature, art, health, among others. In fact, one can see different degrees in which articles can relate to LGBT+. For some topics, LGBT+ is a central element and is visible in its title or first paragraphs (e.g., "LGBT rights in France"). To others, its relation may be less obvious and still appear in a subsection of an image (e.g., in the article "Boston," there are different mentions of the gay parade, including a photo). In Table 1, we can see a list of different types of LGBT+ articles and some examples that we mention in this study. LGBT+ content includes a wide variety of topics, even broader than this shortlist. Still, these are common types of articles: terms, LGBT+ culture, biographies, organizations and people, places, and cultural creations (movies, literature, sculpture, music, etc.).

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Table 1. List of types of LGBT+ articles and examples for each according to their appearance in the study

Type of LGBT+ article	Examples
Terms	LGBT, Homosexuality, Gender dysphoria, Sexual Minority, and Queer.
LGBT+ culture	Fag Hag, Rainbow flag (LGBT), and LGBT culture.
Biographies	David Bowie, Freddie Mercury, Andy Warhol, and Alan Turing,
Organizations and people	European Parliament Intergroup on LGBT Rights, LGBT community, List of LGBT sportspeople, and Arcigay.
Places	LGBT rights in Saudi Arabia, LGBT culture in Paris, LGBT rights in Italy, and LGBT rights in France.
Cultural creations (movies, literature, comics, music, etc.)	LGBT music, Queer Lion, Call Me by Your Name, White Crane Journal, Death in Venice (film), and In Italia Sono Tutti Maschi.

As said, LGBT+ editors are not only interested.[1] For example, Wexelbaum (2019) says that for articles about countries and cities in particular, it is important that there are mentions of LGBT+ rights or events that relate to it. Those articles dealing with scientific or medical information that impacts the LGBT+ community are especially important, and those on public figures and cultures in connection to an LGBT+ identity should also have their information added to their corresponding articles.

Wikipedia categories are a different type of data point employed to express the relationship between an article and a topic. Categorizing articles as related to LGBT is not the same for all topics and in all Wikipedia language editions. For example, the Wikipedia category "LGBT" exists in 94 Wikipedia language editions, which is comparable to the category "Jews" in the number of languages in which it exists (105 Wikipedia language editions), but far from "American people" (143 Wikipedia language editions) or Biology (237 Wikipedia language editions). Wikipedia article categories are useful to classify content. Every Wikipedia language community decides whether to create them or not, which means that, in this case, the rest of Wikipedia language editions have not created the “LGBT” category either because of a lack of will or capacity. Considering that LGBT+ topics are still taboo in many societies, we could think that it is very likely that some articles exist in certain languages but are not labeled as such.

As an example, the article dedicated to the musician David Bowie in the English and Spanish Wikipedia is categorized with using categories related to his music style and albums published, but also as "Bisexual men", "Bisexual musician”, "LGBT musician from England”, and "LGBT songwriters”. In Serbian and Icelandic languages, for example, their versions of the article do not contain any section or links pointing at his sexual orientation—which is a relevant aspect of his public persona as an artist, especially during the 1980s interviews he gave explaining his sexuality (Mirror.co.uk, 2016)—and neither do these two versions of the article belong to any LGBT-related category.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

For biographies, there also exist additional challenges in categorizing them as LGBT+. Some people might not want to have gender and sexuality categorization on Wikipedia because of privacy or concern. On Wikipedia, the Person Task Force is described as "a working group of members of the LGBT studies Wikiproject dedicated to ensuring quality and coverage of biography articles of confirmed LGBT persons" (Wikipedia, "Person Task Force," 2021). Their action is guided by only labeling someone as non-heterosexual when they have come out publicly: "living persons who have come out, and of deceased persons whose sexuality is not in doubt" (Wikipedia contributors, 2021e). A deceased person might be categorized and identified as lesbian, gay, or bisexual if they had documented noteworthy relationships with persons of the same sex or other sexes, such as Marlon Brando.

There exists an LGBT+ Wikiproject for working on Wikidata and introducing LGBT+ data points on its items. Wikidata is a "common source of open data that Wikimedia projects such as Wikipedia can use", and that is especially useful to update Infoboxes in Wikipedia articles automatically. There is a Wikidata Qitem for every Wikipedia article. When two or more languages have an article in common, this relates to a single Qitem (Appendix A provides more details on Qitems). This way, the data points introduced or revised in Wikidata Qitems flow automatically to the different Wikipedia language editions that are connected to it. The LGBT+ Wikiproject calls to add information on the properties sexual orientation (Wikidata property P91), sex or gender (Wikidata property P21), and also to create items about LGBT+ associations (national or local), pride parades, LGBT choruses, bars, film festivals, podcasts, fictional LGBT characters, video games, etcetera.

Research questions

We identified four research questions to explore in this project, which we describe in what follows:

RQ1. What is the existing LGBT+ content, and how is it explicitly characterized as such in the selected Wikipedia language editions?

To measure the LGBT+ content gap, we need to consider the missing articles, that is, an LGBT+ article that exists in one language but not in others, and the missing data points (categories and links) that frame an article and include the LGBT+ points of view. For example, we need to differentiate LGBT+ articles like David Bowie, which is explicitly related to LGBT+ in some languages but not annotated as such in other Wikipedia language editions.

RQ2. What is the share of LGBT+ content in the selected Wikipedia language editions?

Similarly, based on their experience as editors, LGBT+ Wikipedians acknowledge the disparity in coverage of LGBT+ content across language editions. LGBT+ content coverage is a strategic discussion of the conference *Queering Wikipedia* (Grants: Conference/Kawayashu/Queering Wikipedia, 2021). They argue that "while some languages have a good covering of basic LGBT+ related content, others have little to nothing available." We may also wonder if it is a matter of size, in other words, if larger Wikipedia language editions pay more attention to LGBT+ content, or simply they create more content.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

RQ3. What is the visibility of LGBT+ biographies in Wikipedia language editions' featured articles?

"Featured articles are considered to be some of the best articles Wikipedia has to offer, as determined by Wikipedia's editors" (Wikipedia contributors, 2021f). For an article to have such distinction, they need to be reviewed according to accuracy, neutrality, and completeness criteria. As of June 2021, there are 5,927 featured articles in English Wikipedia, 100 of which are related to LGBT+ and have been improved thanks to the coordination and support of the English version of the Wikiproject "LGBT+ studies". Warncke et al. (2015) took the Featured articles from English Wikipedia and observed that LGBT+ biographies were overrepresented among those articles with lower quality articles but with high demand by readers.

RQ4. What is the coincidence between LGBT+ content and local content in the selected Wikipedia language editions?

Miquel-Ribé and Laniado (2018) found that a quarter of the articles of the largest 40 Wikipedia language editions is dedicated to their cultural context. In other words, every Wikipedia language edition contains a considerable amount of content about the territories where the language is spoken, or biographies, traditions, events, and organizations (among other topics) related to those territories. LGBT+ Wikipedians call to create content on topics that could be considered of global interest (e.g., health) but also more localized (e.g., LGBT rights in a specific territory). For example, the Wikipedians Houssem from Tunisia (Knowledge_Equity_Calendar/15/en, 2021) and Bojan from Serbia (Knowledge_Equity_Calendar/1/en, 2021) have organized edit-a-thons in their respective countries to create basic articles around LGBT+ topics, some of which are specifically localized. While Bojan partners with local entities to organize events, Houssem recognizes the difficulties of accessing LGBT+ partners and information (history, arts, etc.). Generally, we can say that even though LGBT+ content is potentially contextual, we do not know how many LGBT+ articles relate to the editors' local context.

In this paper, we answer these questions by measuring different facets of the LGBT+ content in the different Wikipedia language editions. First, we distinguish between biographies of people with an LGBT+ sexual orientation and others that relate to LGBT+ culture (e.g., music, cinema, activism, etc.).

We propose a computational approach to select articles considered as part of LGBT+ content, and we will build upon the existing framework of the Wikipedia Diversity Observatory. This research project addresses the need to measure, characterize, and monitor the coverage of topics using computational methods (Miquel-Ribé & Laniado, 2020).

Once we have obtained the LGBT+ articles, we propose building a simple dashboard tool to retrieve LGBT+ articles from any Wikipedia language edition according to specific features and examine their availability in other language editions. This way, we expect to encourage the exchange and creation of LGBT+ articles across languages.

The rest of the paper is organized as follows. In the following section, we explain the approach to collect the LGBT+ content. We answer the four different questions in four dedicated subsections to availability and categorization, share, visibility, and local content. Next, we present the LGBT+ Gap Tool, which will allow any Wikipedian to look for valuable articles.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Finally, we draw conclusions by mentioning the limitations of the study, explaining the implications for the Wikimedia Movement, and proposing some recommendations.

Methods

In this section, we first describe the methods we employ to collect LGBT+ articles for all Wikipedia language editions. The understanding of this section may require a very basic understanding of the machine learning terminology. See Appendix B for additional explanations on some of those terms used in this section). Our approach builds on top of the existing framework of the project Wikipedia Diversity Observatory (Wikipedia Diversity Observatory, 2021a), which collects, measures, and characterizes the gender and culture gap. The code deployed in Python3 is made available, as well as the resulting databases (marcmiquel/WDO, 2021).

Finally, in the next section (Data and selection of languages), we describe our data source, additional article descriptors used in this study, and the selection of language editions we primarily focus on to answer the research questions.

How to determine if an article is LGBT+?

To answer the RQ1 on what is the existing LGBT+ content and its characterization in each Wikipedia language edition, we want to obtain both (1) a list of LGBT+ articles that can be considered LGBT+ content by only taking into account the data points in the language edition articles, and (2) a longer list with all the existing LGBT+ articles in each language edition, even though they may not contain enough data points within a specific language edition to consider them so.

To generate the selections of LGBT+ articles, we propose a computational approach based on five different steps (see Figure 1). We apply the first three steps of this approach independently in 94 language editions.[2]

1. Generation of a positive ground truth[3] in every language edition using specific Qitems from Wikidata and the occurrence of the "LGBT" keyword in the article title.

2. Extraction of a set of candidate articles with the potential to be considered LGBT+ based on a set of specific data points.

3. Train and apply a machine learning classifier to determine if the candidate articles can be considered LGBT+ or not based on the language edition's positive ground truths, negative sampling, and the values of types of data points of the candidate articles.

4. Merge all the ground truth and then classify articles of all the 94 language editions into a global list of unique LGBT+ articles.

5. Based on this list, we go back to each language edition and select the list of existing LGBT+ articles using the interwiki links that show us the available articles.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

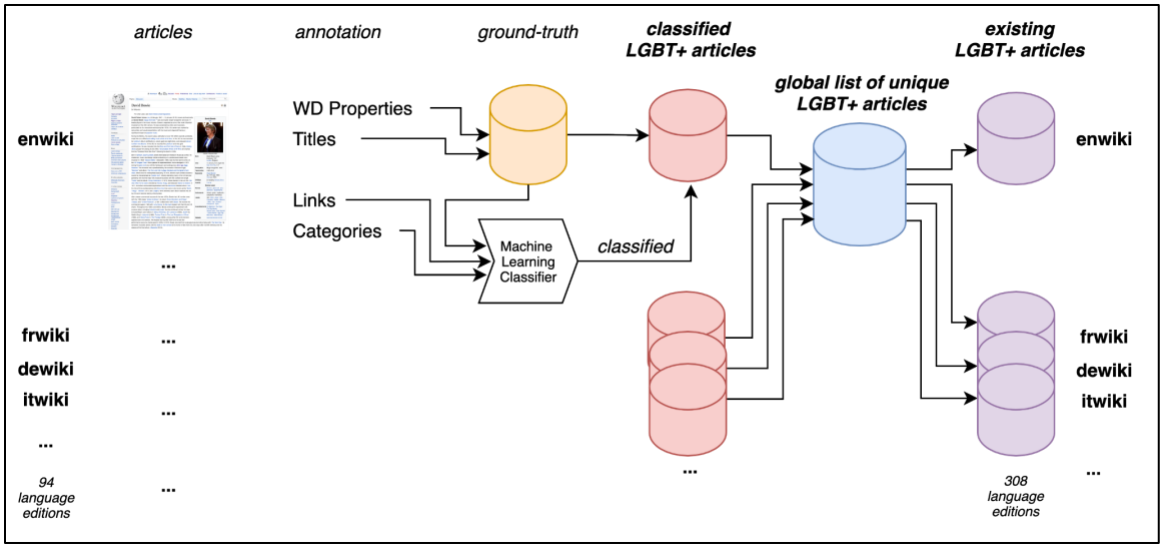

Figure 1. Diagram of the process of multiple steps to obtain LGBT+ articles in every Wikipedia language edition

Data points for the ground truth and the candidate articles

The machine learning classifier takes into account different types of data points listed in Table 2. This approach is similar to the process of collecting "local content" (Miquel-Ribé & Laniado, 2020).

Some data points, when containing a reference to LGBT+, allow us to make a straight decision and classify the article as LGBT+. For example, this would be the case of certain Wikidata properties like sexual orientation or the appearance of some words in the title of an article. Articles obtained through these data points become the ground-truth articles (the positive training set the classifier will use to compare the candidate articles with and evaluate whether they should be part of the LGBT+ Content).

Other data points like the article categories and the article links (i.e., the articles linked in the text of an article) indicate a certain degree of relationship to the topic, but not necessarily conclusive to consider it LGBT+ content. Articles that contain these other data points are candidate articles for the selection of LGBT+ articles.

Step 1: Ground truth articles

Two types of data points allow us to reliably label Wikipedia articles as LGBT+ articles and construct the ground truth: Wikidata properties and the article page titles.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Table 2. List of data points and examples of articles and their values

	Data Points	Description	Example Articles
Ground-truth	Wikidata properties	Articles-Qitems containing WD property sexual orientation (P91) or alternatively properties such as spouse (P26) or partner (P451).	Andy Warhol (Q6636), Alan Turing (Q6636), Drew Barrymore (Q43200).
	Keyword	Articles containing the "LGBT" term in their titles. Value is binary (0 or 1).	LGBT in Islam, List of LGBT rights organizations, LGBT community.
Candidate articles	Category crawling level	Articles whose category is in a category tree whose top category contains the term "LGBT" in its title. Value is the distance from the top.	Korybantes (3), Fag Hag (1), Gender dysphoria.
	Inlinks from (number and percentage)	An article's number of incoming links (inlinks) from LGBT+ articles from the ground-truth and percentage of these inlinks in relation to all the incoming links.	European Parliament Intergroup on LGBT Rights (76, 0.93), LGBT rights in Saudi Arabia (502, 0.807), LGBT community (117, 0.117).
	Outlinks to (number and percentage)	An article's number of outgoing links (outlinks) to LGBT+ articles from the ground-truth and percentage of these outlinks in relation to all the outgoing links in the article's text.	List of LGBT sportspeople (164, 0.220), LGBT music (41, 0.188), LGBT culture in Paris (36, 0.192).

Wikidata properties

As said in the previous section, Wikidata is a shared structured database used by many Wikipedia language editions to centralize some specific facts that are later retrieved and used in Wikipedia articles, as in, for example, the infoboxes.

To follow the same example, the Wikidata Qitem of the singer David Bowie contains information relative to his name, gender, profession, and birthdate, among other aspects. When he passed away, the property "date of death" was introduced on the Wikidata Qitem page. This fact appeared in the David Bowie biography in those Wikipedia language editions whose "infoboxes" are connected to Wikidata.

To examine the relationship between Wikipedia articles and LGBT+, we retrieved the values for three Wikidata properties: P91 (sexual orientation), P26 (spouse), and P451 (unmarried partner). Sexual orientation contains different possible values, including homosexuality, heterosexuality, bisexuality, among others. This is our main approach, and we assume that any value different

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

from heterosexuality in the sexual orientation property value implies that a biography that is LGBT+.

However, since this property is sometimes not used, to increase our selection we have a secondary approach: we use the other two properties: spouse and partner, which indicate the name of the partner when the property's sexual orientation is not used. By examining the gender(s) of the person's partner(s), we assume that the person is homosexual (same gender), bisexual (more than one partner and from different genders), or heterosexual (only partners from the other gender).

Following the same example, David Bowie's sexual orientation property states he was bisexual, even though he had two spouses of the female gender. In this case, sexual orientation is derived directly from the sexual orientation property. But on the contrary, the Russian mathematician Pavel Alexandrov's sexual orientation property is empty; while the spouse and unmarried partner properties are filled with the names of a woman and a man, respectively, therefore it can be assumed that his sexual orientation is bisexuality.

Assuming sexual orientations out of the partners/spouses could be contested, as someone may consider this different from "coming out", since sexual orientations may change over time, and it may be difficult to assume one sexual orientation or another without understanding the personal circumstances. For example, we could mislabel a biography as heterosexual when in fact, it could be that the person had not come out, or as bisexual if the person has come out after having a partner of another gender. Nonetheless, these are very few cases. Another shortcoming of this second approach to obtaining LGBT+ biographies would be the fact that the sexual orientations can only be inferred from binary genders (again, we would only miss a few cases which would not have been already considered as LGBT+ through the sexual orientation property).

However, even though we acknowledge these limitations, we think in terms of cost-benefit, and it is helpful to increase the number of LGBT+ biographies in this way. For example, as of July 2021, the number of LGBT+ articles that have been identified using all three mentioned properties and that exist in the English Wikipedia was a total of 3,235, of which 790 (24.42%) have been identified using the partner/spouse properties, given that the sexual orientation property was empty.

In most languages, the spouse/partners Wikidata values are exposed in the article infobox, and for this reason, we consider that Wikidata values are in-article data points. This, for example, is the case of Freddie Mercury's English Wikipedia article infobox, which includes partners Mary Austin (1970–1976) and Jim Hutton (1985–1991).

Page titles

Article page titles are a concise description of the topic of an article, whether it is a single concept, a relationship between two, or a knowledge domain. Therefore, by looking at article titles and checking whether they contain the keyword "LGBT" in that particular language edition, we can ascertain whether an article belongs to the LGBT+ content. We choose "LGBT" instead of LGBTQ, LGBT+, or any other acronym since LGBT is a common subset of these other terms.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Hence, we first create a dictionary of the term LGBT in as many Wikipedia language editions as possible by taking the “LGBT” article in the English Wikipedia and then obtaining its equivalent title in all the other Wikipedia language editions where there is an equivalent (this is the case for 87 languages). Then, we retrieve all the articles that contain this term in their title, assuming that their content will be closely related to the topic (e.g., in the English Wikipedia, there are articles as varied as “LGBT music," "Rainbow flag (LGBT)," and "LGBT rights in France"). While in principle we could use the same techniques with other terms, we limit it to only LGBT, considering that it is a unique combination that is unlikely to retrieve articles not related to the topic.

Step 2: Candidate articles

Articles retrieved using the above-mentioned Wikidata properties and article titles are reliably part of the LGBT+ content and constitute our ground truth. To expand this selection of articles, we examine the data points “article categories" and the article’s in- and out-links and vectorize them into a vector of five different features, which we later feed into the classifier. The features in detail are:

Feature 1: Category crawling levels

The first feature is derived from the categories which are given to Wikipedia articles. We use the same dictionary we have created for the term "LGBT" in the 87 Wikipedia languages to retrieve a set of categories containing the term in their titles. In the English Wikipedia, with this method we retrieve the category "LGBT" and many other combinations, including "LGBT culture," "LGBT people", etc. Some languages have richer categorization systems than others, being more specific or more structured, and at the same time open and supporting the categorization of LGBT+ content. Others do not even have the main category "LGBT". As of October 2020, when the data was retrieved, there were 87 language editions with this category.

In Wikipedia, articles are categorized according to one or more categories, but categories are also contained in each other, in the form of a treelike graph structure, becoming more and more specialized as we explore the structure level by level. So, in the English Wikipedia, the category "LGBT" contains articles such as "Sexual minority”, but also, other categories such as Queer or “Homosexuality", which in turn contain other categories like "Homosexuality and bisexuality deities". While the treelike structure is generally becoming more specific, it also contains some circular paths and sometimes unconnected categories that have little relationship with the others. By crawling the category graph, we can collect the articles that are directly categorized under the category "LGBT", and also all the other articles at each level of subcategorization. The further from the top level, the less related the articles are to the LGBT+ topic.

For all the articles that were categorized at one level or another, we store the number of jumps from the top level, using it as an indicator of proximity to the LGBT+ topic. Given that the Wikipedia categorization system tends to be wide and articles usually belong to more than one category, we take the shortest number of jumps to the top level, which contains a category with "LGBT" in its title.

Since the categorization is usually exhaustive, this method is useful to obtain almost all the articles that relate to the category title. The first two levels usually include articles that are core to the topic, but sometimes, starting at levels 5-10 from the top, articles become unrelated

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

to LGBT+ due to an unrelated category that has been placed in the category tree. For this reason, one cannot assume that all the articles obtained through the category crawling are LGBT+ content. The category crawling feature is useful as it quantifies the distance between the article topic and that of the category with the "LGBT" term in its title.

Features 2-5: Inlinks from / Outlinks to

The second set of features aims at quantifying articles according to their incoming and outgoing links, starting from the assumption that concepts that relate to an LGBT+ topic are more likely to be linked to one another. We distinguish between inlink-based features (two and three) and outlink-based features (four and five).

Features 2 and 3:

For each article, we counted the number of links (feature two) coming from other articles we already considered as part of LGBT+ content (our ground truth: non-heterosexual biographies and articles containing LGBT+ in their title) and computed the percentage (feature three) in relation to all the incoming links as a proxy for relatedness to LGBT+. The number of incoming links (inlinks) is a clear indicator of relevance because it implies that this article is needed to explain something in the other article which links to it.

Features 4 and 5:

The outgoing links that are placed in the text of Wikipedia articles and point at other articles. Since Wikipedia articles contain links spread over all the text, this collection of links tends to reflect all the different topics relatable to an article relates. Likewise, for each article, we count the number of links (feature four) pointing to other articles that are already qualified as LGBT+ content. An article with a high percentage of outlinks (feature five) to LGBT+ content relates to the topic, and therefore it is potentially a good candidate to be part of that selection. For example, the article "White Crane Journal" from the English Wikipedia is about a gay journal published in San Francisco, and 50% of its outlinks point to other LGBT+ articles.

Step 3: Machine Learning classification

Training and testing

The previously described five features are used to vectorize all the articles from each Wikipedia language edition and to feed a classifier that expands the reliable collection of LGBT+ content. The features used to vectorize the articles are thus: category crawling level, number of outlinks to LGBT, percentage of outlinks to LGBT, number of inlinks from LGBT, and percentage of inlinks from LGBT. The scikit[4] library implementation of the machine learning classifier[5] Random Forest (Pedregosa et al., 2021) is used with 100 estimators in this feature space, following the approach by Miquel-Ribé & Laniado (2020).

To train the classifier, we have the positive ground truth: the articles we consider reliably belonging to the LGBT+ topic. Since we do not have a set of articles of which we know are not LGBT+ content, we employ a negative sampling process (Dyer, 2014). Articles that are not in the positive ground truth are retrieved and introduced in this sampling process. We then use this set to extract five times a set of equal size as the positive ground truth. By using this approach, the

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

classifier is trained to distinguish positive articles from random articles which are not in the ground truth. In Table three, we can see the number of articles in the positive ground truth and in the full training set, which ranges from 1,626 (Serbian) to 21,306 (English). As candidate articles, we use all the articles not belonging to the positive ground truth without excluding any article. The accuracy provided by the classifier is 0.95. The category crawling level in the category tree (feature two) emerged as the most relevant feature.

Manual assessment

In order to evaluate the quality of the selection of LGBT+ articles, two raters perform a manual assessment test in a process similar to the one followed by Miquel-Ribé and Laniado (2018) to evaluate the selection of local content in Wikipedia. For the assessment, we randomly picked 100 articles classified by the algorithm as positive (LGBT+ content) and 100 articles classified as negative (non-LGBT+ content). Each rater manually assigns these articles to be LGBT+ related or not. Then, we compute the F1-score to assess the accuracy of the selection based on the average of the two ratings. The results are presented in Table three, which also details the percentage of false positives (FP) and false negatives (FN) and the precision and recall for each language edition. The assessment finds on average 6.14% false positives and 0.6% false negatives. The average value of F1 is 0.965.

Table 3. List of data points and examples of articles and their values

ISO code	Language	Positive ground truth	Full training set	FP %	FN %	Precision	Recall	F1
ar	Arabic	950	5700	7.5	0	0.92	1	0.961
en	English	3551	21306	10	0	0.9	1	0.947
fr	French	1400	8400	2.5	0	0.975	1	0.987
es	Spanish	1496	8976	1.5	0	0.985	1	0.992
de	German	1407	8442	3.5	0.5	0.96	0.995	0.98
it	Italian	1461	8766	3	0	0.97	1	0.985
ja	Japanese	710	4260	5.5	0	0.945	1	0.972
pl	Polish	961	5766	5.5	0.5	0.945	0.995	0.969
pt	Portuguese	1020	6120	5	0	0.95	1	0.974
ro	Romanian	311	1866	6	1.5	0.94	0.984	0.962
ru	Russian	1090	6540	7.5	2.5	0.925	0.974	0.949
sr	Serbian	271	1626	11	2	0.89	0.978	0.932

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

uk	Ukrainian	655	3930	5	1.5	0.95	0.984	0.967
zh	Chinese	710	4260	12.5	0	0.875	1	0.933

Data and selection of languages

Wikimedia Foundation Dumps

To extract the data points for the articles of each Wikipedia language edition, we employ the dumps (Wikidata, page titles, categories, and links) the Wikimedia Foundation produces on a monthly basis (Wikimedia Foundation, 2021). For the current selection of LGBT+ content, we retrieved those generated in October 2020 for all Wikipedia language editions.

Additional article descriptors

Since this work expands on previous work at the Wikipedia Diversity Observatory (Wikipedia Diversity Observatory, 2021a), we use the available data provided by the project. We use additional article descriptors to characterize the article topic and relevance.

In regards to the topic, we use the “Featured Articles” (Wikipedia contributors, 2021g) descriptor created from the Wikipedia category, a gender descriptor based on the Wikidata property gender, and a “local content” binary which indicates whether an article belongs to the Wikipedia language edition geographical and cultural context (Miquel-Ribé & Laniado, 2019).

In regards to the article relevance descriptors, we take the article creation date, number of Bytes, number of discussions, number of editors, number of edits, number of inlinks, number of inlinks from local content, number of interwiki links, number of outlinks, number of outlinks to local content, number of pageviews, number of references, and the number of Wikidata Properties.

Article topic features are used to answer some research questions, and article relevance descriptors are used in the LGBT+ Gap Tool to rank and filter articles (see Section four).

Selection of languages

In order to answer the research questions, out of the 87 Wikipedia language editions which contain the "LGBT" category, we select a manageable group of Wikipedia language editions that can facilitate the analyses. This group is composed according to two different main criteria: geographical proximity and geographical spread.

By choosing languages from the same geographical context, we want to be able to see if there are noticeable differences between them that we can attribute to their sociocultural factors. In this case, we chose five Eastern European languages: Polish (pl), Romanian (ro), Russian (ru), Serbian (sr), and Ukrainian (uk). However, in order to have robust conclusions on the state of LGBT+ content in Wikipedia, we also want to have languages that are spread and whose Wikipedia language editions have at least 1 million articles. In this case, we chose Arabic (ar), Chinese (zh),

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

English (en), German (de), French (fr), Italian (it), Japanese (jp), Portuguese (pt), and Spanish (es).

Results

In this section, we provide the answer to the four research questions of the study. For each research question, we illustrate our findings with visualizations and describe the results.

RQ1. Existence of LGBT+ content across Wikipedia language editions

With the collection of LGBT+ content described in the methods section, we can answer the first research question (RQ1) on the existence of LGBT+ content in the different Wikipedia language editions. Figure 2A gives an overview of the total amount of LGBT+ content available across all Wikipedia language editions. We find in total 181,250 articles, of which the majority are biographies. Only 41.69% of these articles do not belong to this type. The majority of the biographies (62.6%, 36.5% of all LGBT+ articles) do not have a specific sexual orientation assigned, while 24.8% (14.5% of all LGBT+ articles) correspond to heterosexual orientation and 9.6% (5.6% of all LGBT+ articles) to homosexuality. The remaining 2.9% (1.7% of all LGBT+ articles) are assigned to a bisexual orientation.

While Figure2A counts all the occurrences of an article in multiple language editions, in Figure 2B we count multiple occurrences only once and see the number of distinct articles across all Wikipedia language editions. This list of unique LGBT+ articles contains 43,827 articles or Wikidata Qitems. We find a larger share of articles that are not biographies (46.2%), as well a larger proportion of biography articles (40.85%) with no specific sexual orientation. On the contrary, the proportion of distinct biographies of heterosexual people is much smaller (6.36%), indicating that this category seems to be containing over-proportionally more articles that are more frequently appearing in many language editions. Conversely, the proportion of biographies about persons with a homosexual or bisexual orientation in the global list of unique LGBT+ articles remains quite similar to Figure2A (5.28% and 1.25%, 2,135 and 548 biographies correspondingly). The proportion of other non-heterosexual biographies is small, 0.39% (only 172 articles).

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

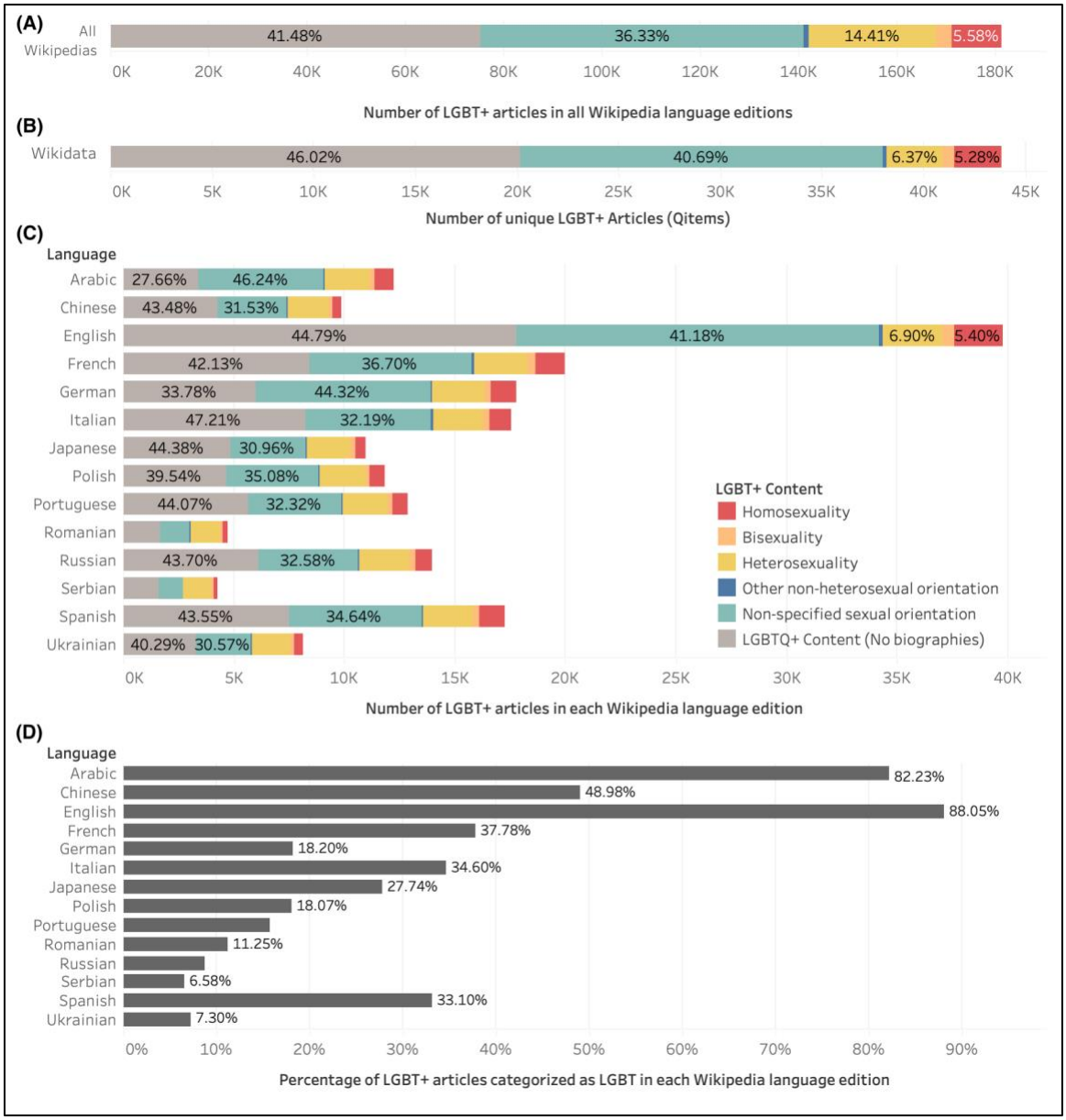

Figure 2. The number of LGBT+ articles for each of the selected Wikipedia language editions

In colour, sexual orientation in case of biography and other LGBT+ content in gray. In (A), aggregated for all the Wikipedia language editions, in (B) for unique Wikidata Qitems, and in (C) for the selected Wikipedia language editions. (D): proportion of ML-classified LGBT+ articles in the 14 Wikipedia language editions in relation to all the existing LGBT+ articles of panel C in those Wikipedia language editions.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Figure 2C shows the number and the composition of the LGBT+ articles in the selected 14 Wikipedia language editions analyzed in more detail in this study. As expected, English is the Wikipedia language edition that contains the largest amount of LGBT+ content (39,759 articles, which corresponds to 90.71% of the number of global unique LGBT+ articles of Figure 2B). In English Wikipedia, 44.79% of the available LGBT+ articles are not biographies. In general, the classification algorithm collected many biographies of heterosexual and people of undetermined sexual orientation who may have been an activist or supporter of LGBT rights (between 30 and 46% depending on the language edition), and in all cases, non-heterosexual biographies correspond to between 6 and 10% of all content that may be related to LGBT+. The proportions between LGBT+ biographies are quite similar across languages regardless of their total number of LGBT+ articles (homosexuality is between 4.31% and 6.42%, bisexuality between 1.31% and 2.22%, and other non-heterosexual orientations between 0.40% and 0.71%).

Finally, Figure 2D shows the actual percentage of the articles that are ML-classified as LGBT+ articles thanks to the data points in the 14 Wikipedia language editions (i.e., article titles, article categories and links) in relation to all the existing LGBT+ articles in those Wikipedia language editions shown in 2C. English, Arabic, Chinese, French, Italian and Spanish in descending order are the languages with the largest proportion being classified by the ML-algorithm, having all of them a proportion larger than 30%. This means that in these languages it is more common to provide LGBT+ information in the articles which have some relation to LGBT+. In addition to having more LGBT+ articles—from the global selection of unique LGBT+ articles—than the others, English Wikipedia also stands out as the language, which has more LGBT+ articles that have been classified as such thanks to the data points in relation to them.

RQ2. Share of LGBT+ content in Wikipedia language editions

The second research question (RQ2) inquiries on the share of LGBT+ articles. When analyzing the share of LGBT+ content in the selected Wikipedia language editions, we first look at the share of biographies with sexual orientation data points that is available in the 14 selected Wikipedia language editions (shown in Figure 3A). The proportion lies between 17.79% for the biographies in the Romanian Wikipedia and 4.18% in the English Wikipedia, which is, in general, a low percentage given the relevance of the characteristic. Furthermore, this percentage includes those Qitems with the sexual orientation property or with spouse/partner properties. This percentage is low, especially if we compare it with the proportion of biographies with a gender assigned, which stands above the 99% of biography Qitems in Wikidata.

The fact that the English Wikipedia has the lowest percentage, but still in absolute numbers the highest number of biographies with sexual orientation data points, while on the contrary Serbian and Romanian have the lowest absolute numbers but the highest percentage, might mean that the editors of the English Wikipedia create more biographies but do not populate the corresponding information in the Wikidata properties at the same rate. In contrast, in the case of the other two language editions, they create them when they already exist in other languages and therefore are already more complete in Wikidata.

In Figure 3B, we can observe the proportion of non-heterosexual orientations among the biographies with sexual orientation data points. The values lie between 4.4% for Arabic as the highest and 2.9% for Japanese as the lowest. Interestingly, this 3.5% in the English Wikipedia is in agreement with the percentage of people in the US who self-identify as LGBT (Gates, 2011; Gates, 2017). It seems, however, that the share of bisexuals is underrepresented in relation to

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

the homosexual orientation, at least in relation to self-identification (where roughly the same amount identifies as either homo- or bisexual). Although, we should state here that it is difficult to compare these numbers directly to statistics of currently living people, as biographies involve many people from periods in history where non-heterosexual orientations were less socially accepted and thus not acknowledged publicly in many cases.

At the same time, we should mention that comparing the proportion of non-heterosexual biographies to the total number of biographies with sexual orientation data points might not reflect the proportion of LGBT+ biographies, given that the percentage of usage of the sexual orientation data points is low. It could well be that sexual orientation data points (i.e., properties) are not introduced because heterosexuality is considered the expected answer. At the same time, the interest in having this information public is higher by part of the LGBT+ community, who wants to give visibility to all people who have already "come out".[6] For this reason, we estimate that the real percentage of LGBT+ biographies might be somewhere between the percentages shown in Figure 3B and those of Figure 3C, which are computed in relation to all the biographies. There, we appreciate that the range of non-heterosexual orientation biographies is between 0.18% in English and 0.61% in the Serbian Wikipedia.

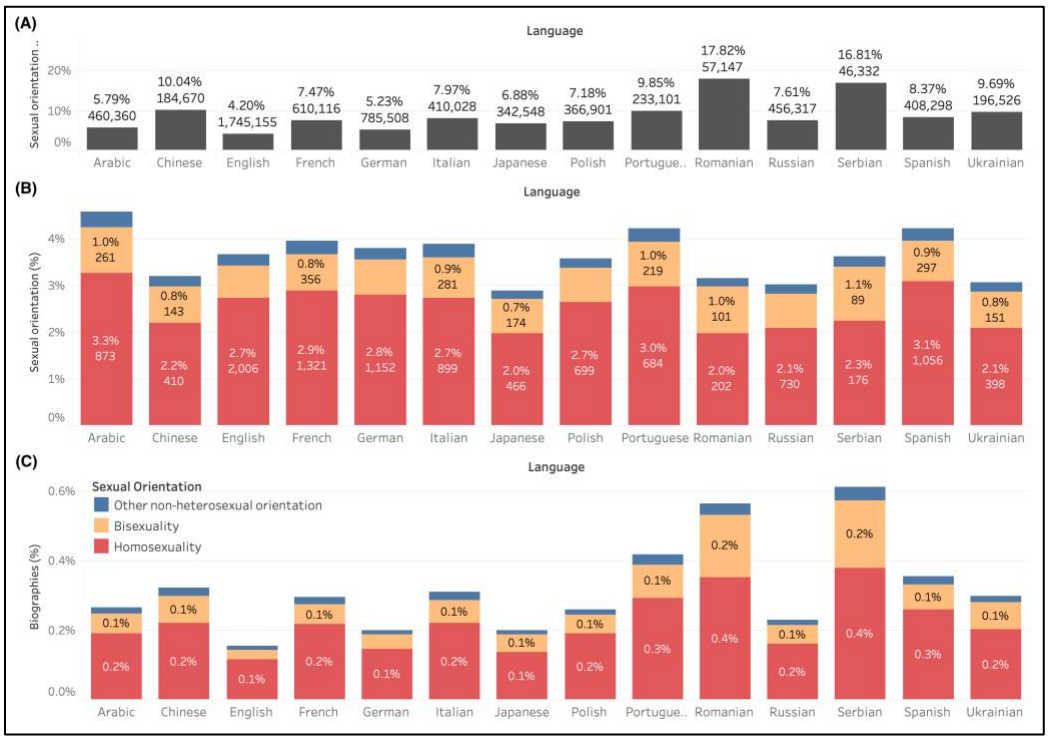

Figure 3. Biographies and sexual orientation. (A): number and percentage of biographies with sexual orientation-related properties (Wikidata) for the selected Wikipedia language edition. (B): percentage of biographies with sexual orientation-related properties of homosexuality, bisexuality, and other non-heterosexual orientations among all the biographies with sexual orientation-related properties. (C): the percentage of non-heterosexual biographies calculated with respect to the total number of biographies.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Finally, in Figure 4, we analyze the actual share of LGBT+ articles in the selected language editions and obtain an answer to RQ2. The figure compares this share (in %, x-axis) with the total number of articles in the selected Wikipedia language edition (y-axis). We observe no clear relation between the two quantities. Portuguese is the language where it has the largest share (1.4% of all its articles contain LGBT+ content), followed by Romanian and Arabic, Spanish and Italian.

On the lower end, we find German, Serbian and English. The latter has a share of only 0.71%. This may be surprising, given that English has the largest number of LGBT+ related articles (as shown in Figure 2). However, since the English Wikipedia already covers nearly all distinct LGBT+ articles, there seems to be little room for improvement left. This limitation, however, is not the case, for example, in the German or French Wikipedia, which cover 40.57% and 45.56% of all the global unique LGBT+ articles (shown in Figure 2C), and whose share in relation to their total number of articles is 0.71% and 0.88%.

We can generalize that covering more LGBT+ articles does not correlate with a higher share. In fact, one might erroneously assume that a higher share is indicative of more interest in the topic. However, even though there exist important efforts by groups of LGBT+ Wikimedians creating content or introducing references to the topic, the overall creation of articles in Wikipedia language editions happen in general in a distributed and spontaneous fashion according to the interests of different profiles of editors. For this reason, from the Wikimedians point of view, it seems more valuable to use the proportion of coverage or the overall number of existing LGBT+ articles in a language edition (as shown in Figure 2) to track progress towards the goal of increasing LGBT+ information in each language edition.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

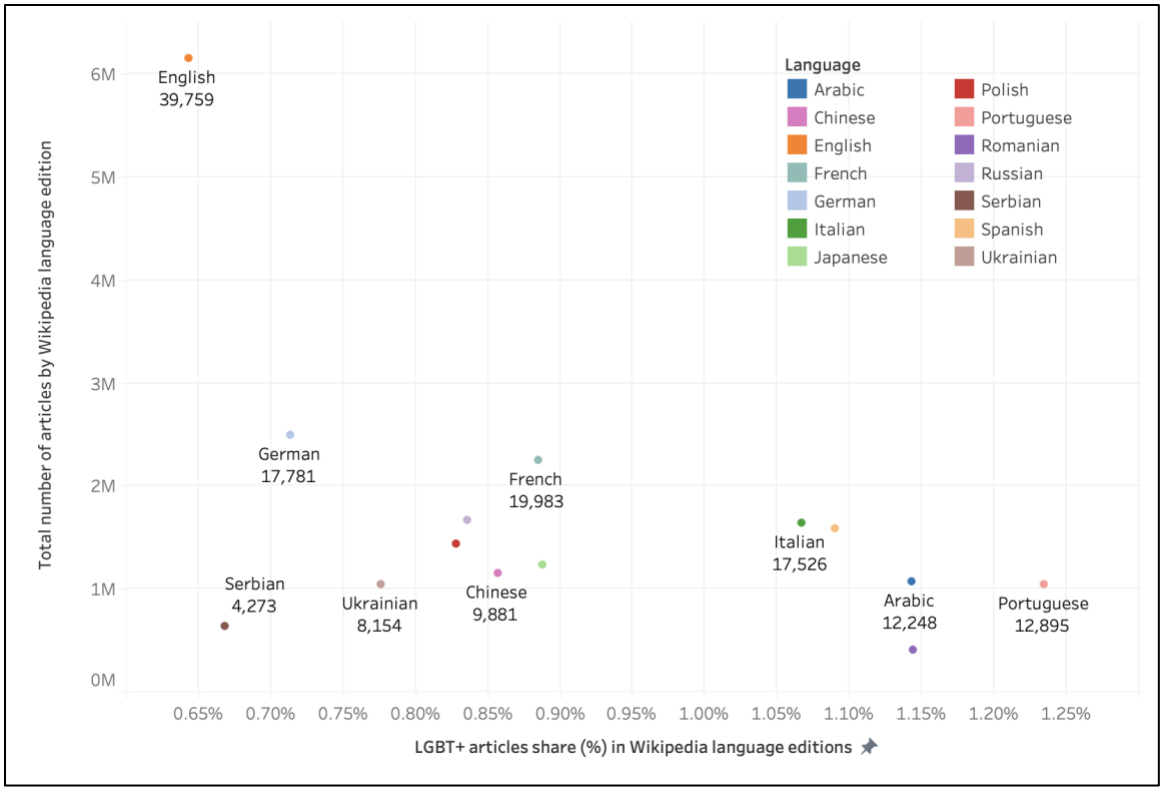

Figure 4. Absolute Number vs Share of LGBT+ articles. (y-axis) the number of LGBT+ articles in the selected Wikipedia language editions, and (x-axis) the percentage of share according to their total number of articles. Below the name of the Wikipedia language edition, we can see the number of existing LGBT+ articles.

RQ3. Visibility of LGBT+ biographies in Featured Articles

The third research question (RQ3) inquiries about the visibility of LGBT+ articles, more precisely of LGBT+ biographies. To study it, we observe their proportion among the biographies which are included in the category "Featured Articles". We analyze these biographies by sexual orientation and gender. This is depicted in Figure 5A, which gives an inconclusive picture. While for some Wikipedia language editions, the share of non-heterosexuality in "Featured Articles" biographies is larger than if all biographies are considered (as had been done in Figure 3), other languages do not have a single featured biography with a non-heterosexual orientation. In particular, this is observed for Chinese, Japanese, and Serbian. However, it should also be stated that the total number of featured biographies in these language editions is very small, in the order of 10 or less.

In regard to the gender distribution, we first observe that as of the moment this article was written, there only exist biographies with one of the two binary genders in Featured Articles. In English Wikipedia, we can see for both males and females that the proportion of non-heterosexual biographies in Featured Articles is larger than the proportion of non-heterosexual biographies taking into account all biographies. Non-heterosexual males have a share of 9.5% of

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

featured biographies, which increases to 15.2% if we include the ones with a not-specified sexual orientation. For females, these shares are a bit smaller, with 8.6% and 13.9%, respectively. Instead, for other languages like German, the non-heterosexual females have a greater percentage than non-heterosexual males (25% with respect to 10%). The shares are also exceptionally high in the Spanish and Portuguese language editions.

In Figure 5B, we see the number of articles for males and for females in featured biographies, which shows that the gender gap is also present in this group of articles. If we compare it with current gender gap data from Humaniki's dashboard (Humaniki | Wikimedia Diversity Dashboard Tool, 2021), we see that, for example, in English Wikipedia, there exist only 18.94% female biographies, which relaxes to 37.91% in featured biographies. The same happens in nine other languages: Arabic (16.22% and 24.14%), Chinese (18.98% and 33.33%), German (16.55% to 19.35%), Polish (16.65% and 34.78%), Portuguese (19.00% and 31.37%), Romanian (18.45% and 58.82%), Serbian (18.98% and 71.43%), and Ukrainian (16.70% and 33.33%). Romanian and Serbian even reverse the gender balance and go beyond parity in featured articles. On the contrary, the French and Italian editions have the biggest imbalance in the number and proportion of featured biographies (only 6.45% and 9.09%, compared to 18.82% and 15.96% among the total number of biographies). Those two language editions have a noteworthy proportion of non-heterosexual biographies, only among male biographies.

Another interesting gender-related trend we observe is that the share of bisexual featured biographies is higher for females than males in seven languages out of eight which feature at least one bisexual biography (the exception is the French Wikipedia). In most cases, the proportion of female bisexual biographies is several times higher than male's (e.g., 16.67% bisexual female in German for 4% of bisexual males, or 5.17% bisexual female in English for 2.11% of bisexual male).

While it is hard to find a rationale to explain why some genders or languages give more visibility to non-heterosexual biographies in their featured articles, we must acknowledge that this compensates for the low proportion of biographies overall we have seen in the previous figures. In this sense, working on featured biographies seems a more attainable goal for an organized group of Wikimedians like Wikimedia LGBT+, focusing on quality rather than plain quantity.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

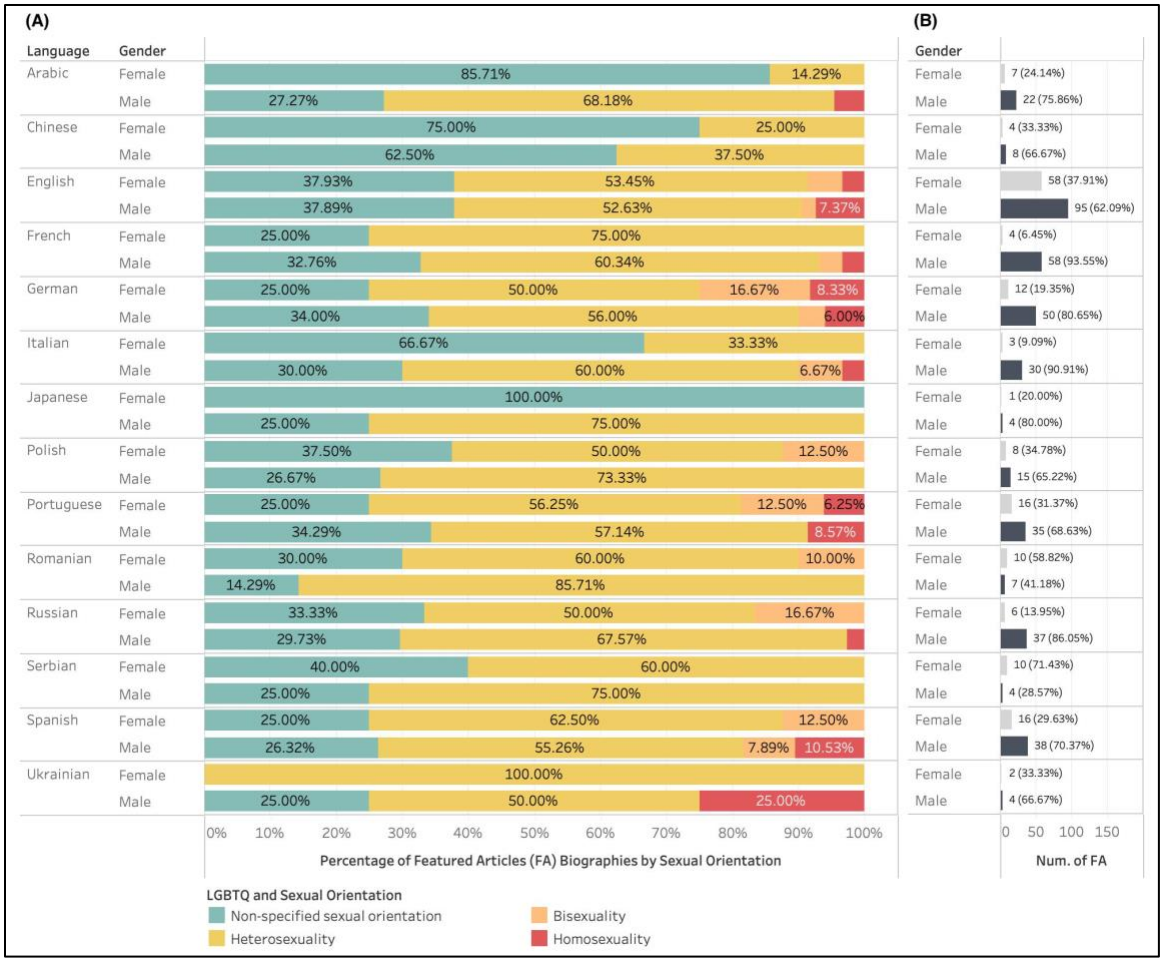

Figure 5. Featured Articles (FA) Biographies by Sexual Orientation. (A), percentage of Featured articles "biographies" by sexual orientation for the available genders. (B), number and percentage of featured article biographies by gender.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

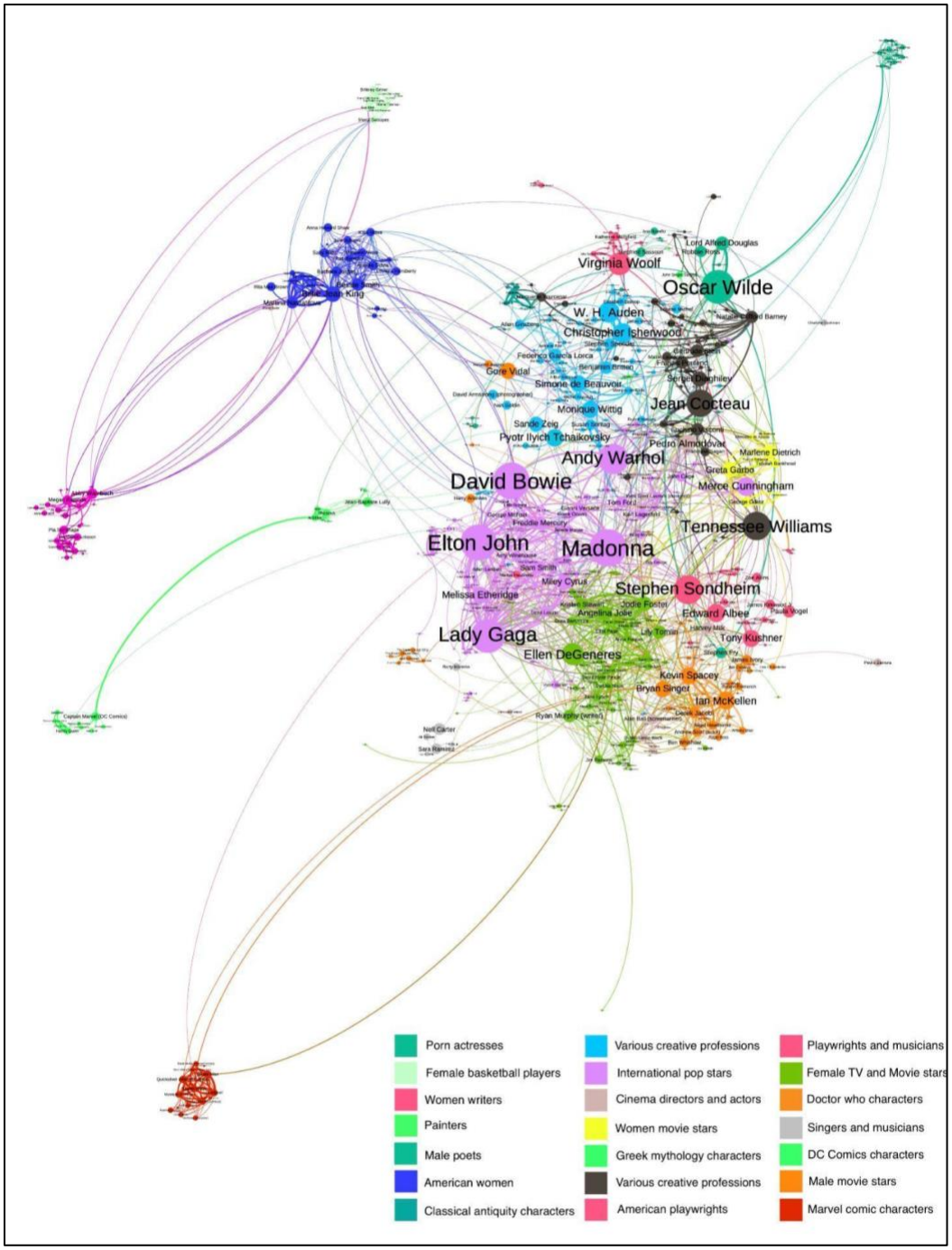

Figure 6. Network graph generated with the links between the Wikipedia articles that are LGBT+ biographies from the 14 Wikipedia language editions.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

In Figure 6, links are only drawn if they exist in at least 2 of these language editions. Following a standard convention in graph representation, edges are drawn curved clockwise to indicate directionality. Colours are assigned according to the clusters identified by an automatic clustering algorithm (the Louvain method) to highlight groups of LGBT+ biographies that are more connected between each other. Node size is proportional to the page rank of the biography in the network. Only the giant connected component is shown.

Finally, to navigate another aspect of the visibility of the LGBT+ biographies in the Wikipedia language editions, we considered the number of incoming links for each biography in each languages' existing LGBT+ articles. Figure 6 depicts the giant connected component of the resulting network, considering only links that appear in at least 2 of the 14 analyzed language editions. This network has been created following a standard convention in graph representation. Edges are curved and drawn in the clockwise direction to indicate directionality. Colours are assigned according to the clusters identified by an automatic clustering algorithm, the Louvain method (Blondel et al., 2008), to highlight groups of LGBT+ biographies that are more connected to each other. Node size indicates the importance (centrality) of the biography in the network, measured through its page rank value.

By taking a quick look at this figure, we can see that the network is dominated by artists and writers mostly from the Anglo-Saxon cultural sphere of influence. Still, it also shows some interesting communities of athletes. Less expected clusters are some fictional characters from the universes of Marvel and DC Comics, a group of porn actresses, and Greek mythological figures. On the bottom right of the figure, we can see the entire legend listing all the different clusters we manually identified and named according to the profession or background of the most prominent nodes in each cluster.

RQ4. Coincidence between LGBT+ content and local content

The fourth and last research question (RQ4) inquiries on the degree of coincidence between the existing LGBT+ articles in each Wikipedia language edition and their share of articles that are considered "local content". A previous study by Miquel-Ribé and Laniado (2018) found that "local content" (denominated Cultural Context Content by the authors) is a considerable proportion of all articles in each language edition, encompassing a wide variety of topics especially related to geography, people and language.

In order to link our results with the previous study, we analyze the share of LGBT+ articles in each language edition that is also part of their share of "local content". Given that an important part of the LGBT+ content is contextual (e.g., people, rights, events, etc.), we expected an important coincidence. Figure 7 depicts the corresponding results. Its top panel (Figure 7A) shows in dark grey the percentage of LGBT+ content that relates to the language's "local content", which is usually below 10%.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

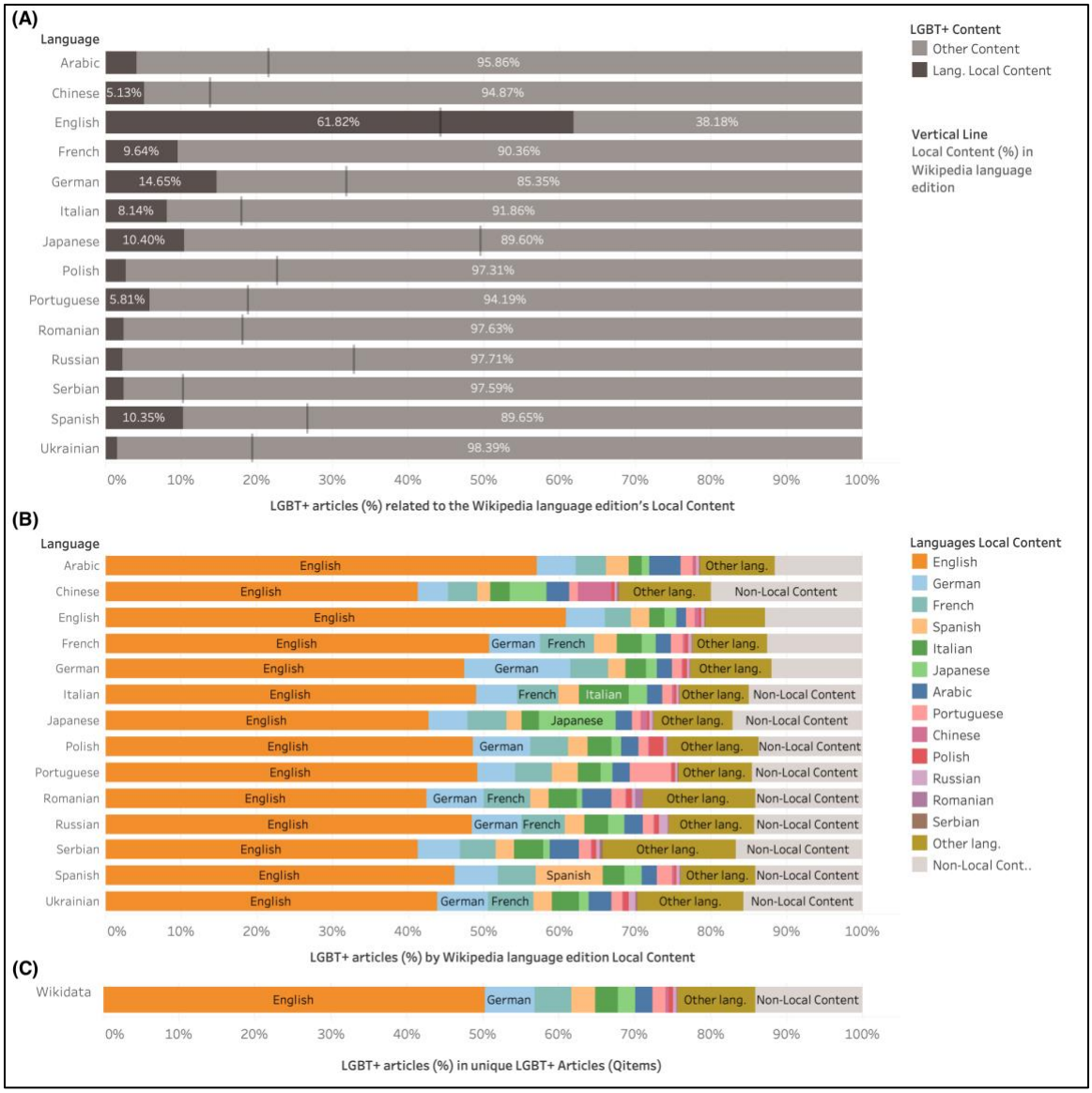

Figure 7. LGBT+ content and Local Content overlap. (A): share of the Wikipedia language edition's "local content" (in dark grey) among the LGBT+ selection articles. Vertical dark line shows the overall percentage of local content in the corresponding Wikipedia language edition. (B): percentage of each Wikipedia language's LGBT+ content which is local content from a specific Wikipedia language edition. (C): percentage of the list of unique LGBT+ articles (Wikidata Qitems) which is local content from a specific Wikipedia language edition.

The vertical dark line depicts the percentage of local content computed in relation to the entire Wikipedia language edition's articles. This shows that the percentage of "local content" among the LGBT+ articles is lower than in the whole Wikipedia language edition except for the English

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Wikipedia (where we find 61.82% LGBT+ local content with respect to 44.25% general share of local content).

On the one hand, the second language edition whose LGBT+ content has a substantial share of "local content" is the German Wikipedia (14.65%, compared to 31.84% of "local content" in the entire German language edition). On the other hand, the absolute distance between the shares of LGBT+ "local content" and "local content" in general in the Japanese Wikipedia is the largest (10.40% vs 49.56%). For Russian, Romanian, Serbian, Ukrainian, and Polish, the LGBT+ "local content" share is as low as approximately 3%, which probably indicates that many LGBT+ articles about rights, activists, movies, books, and so forth, local to these countries or language communities, may not exist yet. The proportion of "local content" among every Wikipedia language edition's LGBT+ articles might be larger if we take into account the LGBT+ articles classified in each language edition rather than all the existing LGBT+ articles. In Figure 7B, we enrich the previous analysis and show the percentage of each Wikipedia language's LGBT+ content that is local content to a specific language edition. We observe that English LGBT+ local content (orange bars) is a very important part of all the other language edition's LGBT+ content analyzed here. German (light blue) and French (turquoise) local content also takes an important proportion of the other language edition's LGBT+ content. The part that does not relate to any of the 14 languages' local content corresponds to only between 20% and 30% of the content (depicted as "Other lang" in brown, but as well includes in gray content that cannot be considered local in any specific language edition). However, in other language editions, the share of their local content is considerably smaller, which means there is a margin for growing their LGBT+ content by creating more articles that relate to their most immediate geographical and cultural background. Finally, in Figure 7C, we depict the percentage of unique LGBT+ articles (Wikidata Qitems) which are local content from a specific Wikipedia language edition. While English LGBT+ local content makes up 50.31% of the unique LGBT+ articles, the distribution is very similar to the ones we saw in the Wikipedia language editions.

Finally, in Figure 8, one can see two network graphs analogous to the one presented in Figure 6. In this case, rather than taking all the biographies from the 14 selected languages, we focused on the biographies of the Spanish (panel A) and French (panel B) LGBT+ local content. The graphs have been generated using the same convention, but using only the articles that exist in this language edition. Figure 8A shows the clusters from Spanish local LGBT+ biographies, which includes the contexts relative to all the South American Spanish-speaking countries as well as Spain. On a first look, we can see that most clusters are dominated by actors and activists, followed by writers, politicians and designers.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

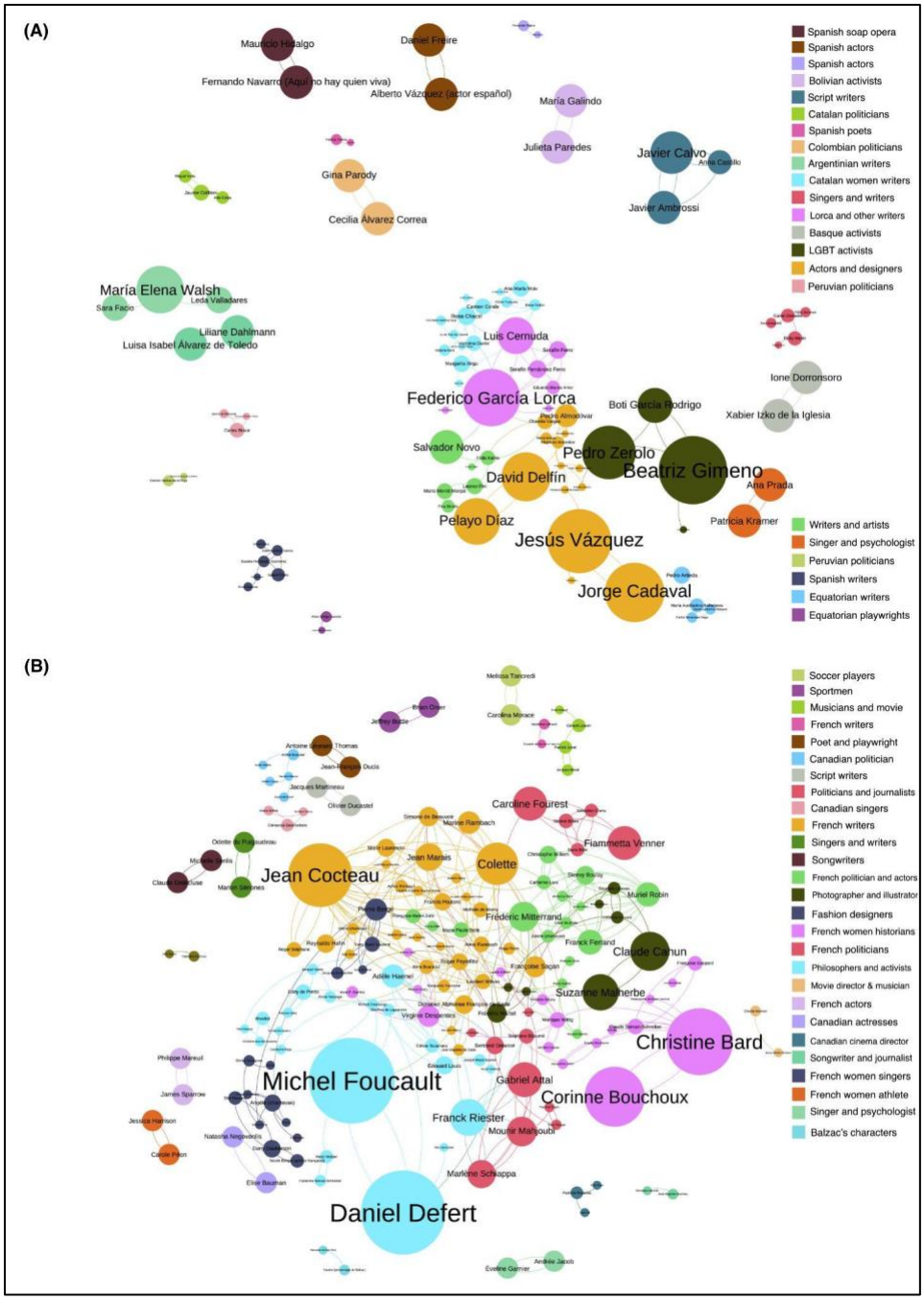

Figure 8. Network graphs generated with the links between the Wikipedia articles that are local LGBT+ biographies in Spanish (panel A) and French Wikipedia (panel B)

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

In Figure 8B, we can see the graph generated for the French LGBT+ local content, which is mainly spread over France and Canada, with no public figures from the rest of the Francophonie - this includes all the countries where French is lingua franca and is especially present in Africa. On a first look at the graph, we see the influence of thinkers like Michel Foucault and Daniel Defert (turquoise cluster) and writers like Jean Cocteau and Colette (orange). While the French LGBT+ biographies also include actors, activists, and politicians, we see a wider variety of profiles in the larger number of clusters.

In general, these biographies are related to profiles in which public appearances are important and reinforce their professional opportunities, whether they are in writing, performative arts, acting, or politics. The graphs reflect the social nature of these professions, with groups of professionals who share some traits, styles, or influence one another. One might think that these are LGBT+ related professions, but in fact, a Wikidata query on the number of Qitems by profession (property P106) shows that these professions are very common for biographies in general.

LGBT+ Gap Tool

In this section, we address the research objective of building a simple tool to assist in bridging the LGBT+ content gaps, which we named "LGBT+ articles dashboard" (Wikipedia Diversity Observatory, 2021b) and is hosted along with other dashboards of the Wikipedia Diversity Observatory.

As stated in the introduction, the main requirement for the tool is to retrieve LGBT+ articles from any Wikipedia language edition filtered and ranked according to specific features, and to indicate their existence in other language editions. Therefore, it shows valuable articles and encourages editors to take immediate action and bridge the gaps when these are not available in other Wikipedia language editions. In this sense, it is similar to other tools like the Gap Finder (Wulczyn et al., 2016), but focused on LGBT+.

The tool allows through a simple graphical interface to choose one "source language" to retrieve articles from one or more "target languages" to verify the existence of the corresponding equivalents to the Qitems of these retrieved articles in them. For example, in Figure 9 we can see the resulting list of articles from a query on the Italian Wikipedia with their titles in the third column and in the *Target langs.* column, we see the availability in a few selected languages for this specific query (English, Romanian, Japanese, and French) with their langcode. The right-most column shows the title in the first selected Target language, in this case, in English.

In order to prioritize among the 17,526 existing LGBT+ articles in Italian Wikipedia, we needed to set some criteria to limit this scope by filtering and ranking them. In this case, the articles were retrieved with a filter that limits the results to "local content without biographies". This filter can be selected at query time in a dropdown list of different topics. Because of this intersection between "local content without biographies" and LGBT+, we see many cultural creations in the list titles, including movies like "Call Me by Your Name (film)" and "Death in Venice (film)", graphic novels like "In Italia sono tutti maschi", and more contextual articles like "LGBT rights in Italy" and "Homosexuality in ancient Rome". While all these exist in the English Wikipedia, we see notable gaps in Japanese and Romanian.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

In fact, to see the gaps more clearly, we can use the dropdown menu "Show the gaps" and select one of the options "Only gaps", "At least one gap", or "No language gaps". The first one is useful in order to find articles missing in all the selected target languages. However, the most usual case may be using a single target language, since Wikipedians most usually focus their activity on one project only. The other two options may give a sense of which articles are partially missing and totally covered.

To rank the results, we employed a newly created feature ("LGBT+ Indicator"), which counts the number of language editions an article has been classified as LGBT+ in. This is a proxy for the interlanguage agreement on how much an article is perceived as belonging to the topic— being its value one when it is only selected as LGBT+ in one language edition, and 94 at maximum (all the languages in which there was the "LGBT+" category).

In the table, we see that the first result is the film "Call Me by Your Name," which has an LGBT+ indicator of value 12, and it has 45 Interwiki links, which means that it exists in this number of language editions. The distance between these two numbers implies that the data points (the categories or the links to and from it) in the article versions of 33 language editions were insufficient for the ML-classifier to label it as LGBT+ content. The value of the LGBT+ indicator is a clear absolute reference to find a valuable gap because it implies that the article contains references to the LGBT+ topic in many languages.

LGBT+ articles retrieved from Italian Wikipedia and its coverage by the target languages

Nº	Qitem	Italian Title	Editors	Edits	Pageviews	Interwiki	Bytes	Creation Date	LGBT Indicator	Target Langs.	English Title
1	Q25136757	Chiamami col tuo nome (film)	69	511	2909	45	105.7k	2017-01-22	12	en, ro, ja, fr	Call Me by Your Name (film)
2	Q1286907	Diritti LGBT in Italia	66	408	324	12	74.3k	2010-09-09	10	en, fr	LGBT rights in Italy
3	Q117546	Omosessualità nell'Antica Roma	69	347	285	13	120.0k	2007-02-15	10	en	Homosexuality in ancient Rome
4	Q742308	Morte a Venezia (film)	62	181	423	32	14.8k	2009-01-24	8	en, ja, fr	Death in Venice (film)
5	Q914157	Teorema (film)	57	106	205	22	5.4k	2005-11-19	7	en, ja, fr	Teorema
6	Q637369	Arcigay	131	550	90	11	30.4k	2004-12-28	7	en, fr	Arcigay
7	Q2504862	Queer Lion	25	113	10	17	7.3k	2007-08-25	6	en, ja, fr	Queer Lion
8	Q676694	Improvvisamente l'inverno scorso	35	68	7	4	3.9k	2008-07-28	4	ja, fr	
9	Q1712162	La maschera di scimmia	25	35	22	7	9.6k	2010-12-24	4	en, ja, fr	The Monkey's Mask
10	Q3797119	In Italia sono tutti maschi	19	29	13	4	2.7k	2008-09-30	4	en, fr	In Italia Sono Tutti Maschi
11	Q1158870	Ferraù	33	102	34	3	6.4k	2007-12-26	4	en	Ferragut
12	Q464384	Tomba del Tuffatore	63	193	333	9	15.1k	2007-07-17	4	en, ja, fr	Tomb of the Diver
13	Q921750	Femminiello	43	106	245	8	7.3k	2008-01-28	3	en, fr	Femminiello
14	Q53572230	Euforia (film)	11	22	679	6	3.7k	2018-05-15	3	en, ja, fr	Euphoria (2018 film)
15	Q16590824	Pisa79	18	58	1	2	13.1k	2014-01-21	3		

Figure 9. Results from the Wikipedia Diversity Observatory dashboard "LGBT+ Articles" (Wikipedia Diversity Observatory, 2021b) for a query to Italian Wikipedia LGBT+ articles and their availability in a few selected languages for this specific query (English, Romanian, Japanese, and French)

Articles in Figure 9 are filtered by "local content without biographies" and ranked in descending order by the number of languages in which they are classified as LGBT+ (LGBT Indicator column).

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Middle columns show some relevant features (e.g., Editors, Edits, Pageviews, Interwiki links, Size in Bytes, and creation date). The right-most column shows the title in the first selected Target language (in this case, English).

In addition to the LGBT+ indicator, the tool also allows using other features to sort the articles and adds them to the table of results as additional columns. In Figure 9, we can see six features that explain very different aspects that characterize a content gap: engagement (number of editors and number of edits), popularity (number of page views), multilingual spread (number of interwiki links), length (number of Bytes), and time (creation date). Additional features can also be found in the dropdown menu "Order by feature".[7]

All in all, having lists of valuable articles about any topic is a very Wikipedian way to identify and coordinate to fill content gaps. "Vital articles" (Wikipedia contributors, 2021h), "List of articles every Wikipedia should have", ("List of articles every Wikipedia should have," 2021), and "Wiki99" ("Wiki99," 2021) are examples of this consensus-based approach with different levels of popularity across Wikipedia language editions. The Wikimedia LGBT+ affiliate has prepared one Wiki99 for LGBT+ topics, including articles on concepts, violence, sex and health, activism, and biographies. We believe that by using the LGBT+ articles dashboard, editors will be able to expand their lists, filtering with some topics and ranking them by the different available features.

Conclusions

While social media has been useful to the LGBT+ community to self-express their identities and promote activism, Wikipedia provides the opportunity to gather all the relevant LGBT+ information that any person might need in more than 309 language editions that are spread over the entire globe. In order to create the articles, Wikipedians access all kinds of online sources and at the same time partner and engage with GLAM professionals to help them share the information.

In fact, creating articles or doing simple edits can be relatively easy, but covering entire topics requires more structured work. To this purpose, Wikipedians create Wikiprojects to list pending articles and organize edit-a-thons to devirtualize and create articles, and very often close partnership deals to incorporate a public institution database. In this sense, Wikimedia affiliates are an essential infrastructure to work at this strategic level and cover those content gaps that are more difficult to fill with spontaneous edits.

The LGBT+ community is spread over multiple languages and spaces with different levels of organization and engagement at both local and global scales, and even though there is an explicit recognition that metrics are necessary for strategic reasons, there are currently none available. To fill this need, we have presented a computational approach to collect LGBT+ articles in Wikipedia language editions along with four research questions to understand the nature of the LGBT+ content gap. To answer them, we selected 14 language editions to study LGBT+ articles and their coverage, share, visibility in Featured Articles, and overlap with language editions' local content.

The research conducted in this paper builds on the previous literature on measuring and monitoring content gaps (Miquel-Ribé & Laniado, 2020; Redi et al., 2020) by specifically addressing the LGBT+ content gap. Its insights contribute to a better understanding of the

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

currently available LGBT+ content in Wikipedia language editions, which we believe is necessary to both create tools that allow regular monitoring of the gaps as well as to design and improve the strategies for content creation. In the following subsections, we will highlight the main findings of this research study, we will explain the limitations of our approach as well as future lines of research, and finally, we will make some recommendations to the Wikimedia LGBT+ community and its collaborators in the GLAM.

Bridging LGBT+ Content Gap

To answer the first research question (RQ1) about the existence of LGBT+ content, we collected a global list of unique LGBT+ articles by generating first a set of LGBT+ articles for each Wikipedia language edition with machine learning algorithms. Then we merged these sets and examined the availability of this global list in each language edition. For the machine learning classifiers, we employed features derived from Wikidata properties (sexual orientation and partners), article titles, Wikipedia categories, and article links. The result showed that as of October 2020, a considerable part of the LGBT+ content (43,827 distinct articles) exists across Wikipedia language editions, being covered best by the English Wikipedia. The LGBT+ articles that exist in each language edition are considerably more than those the classifiers actually classify as such. This means that for a given language, there are many articles whose versions in other languages allow them to be classified as LGBT+ articles but do not form part of any LGBT-related category or do not contain sufficient links to other LGBT+ articles in this specific language version.

An examination of the share of LGBT+ content in each language edition (RQ2) shows us that even though the list of existing LGBT+ articles contains a wide variety of subtopics, it only accounts for 0.5 to 1% of all the content in the 14 examined Wikipedia language editions. For a Wikipedia language edition, covering more LGBT+ articles does not correlate with it having a higher share among all the articles in that language edition. We also found that the share of LGBT+ biographies is around 4% of all the biographies with sexual orientation property in Wikidata and 0.5% of all biographies.

By taking a look at biographies with different non-heterosexual orientations, we saw that even though they are a small share of all the biographies, they are especially visible in the Featured Articles. In several language editions, the proportion of non-heterosexual biographies in Featured Articles is larger than the proportion of non-heterosexual biographies when we take into account all the biographies or those with sexual orientation property in Wikidata (RQ3). Among Featured Articles biographies, the proportion of homosexual biographies is larger in males than in females, and the proportion of bisexual biographies is actually larger in females than in males. Leaving Featured biographies aside, we took all the LGBT+ biographies in all languages and looked at how they link to one another in order to find tighter connected subgroups and those biographies more central in the resulting network. We found Anglo-Saxon pop stars, writers, and film actors to be among the most prominent.

An analysis of the coincidence between each Wikipedia language edition's local content and LGBT+ articles has revealed that in general, the proportion of local content among LGBT+ content is lower than among the entire number of articles of the Wikipedia language editions (RQ4). These are surprising results given that many of the LGBT+ articles are related to geography, social events, or biographies. However, the only exception of a language with a higher amount of local content among LGBT+ articles was English (61.82% with respect to 44.25% of share among

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

the entire Wikipedia language edition). Results allow us to suggest the existence of a considerable margin to create local LGBT+ content in the other language editions.

Lastly, we created a dashboard tool to help in the retrieval and identification of LGBT+ content gaps ("LGBT+ articles dashboard"). Even though it is still at an initial stage (Alpha development), it allows already searching for LGBT+ articles in one language edition according to different criteria. With this tool, we addressed the research objective of providing immediate steps to bridge the gaps.

Limitations and future lines of work

Through the selection of articles that compose LGBT+ content, we investigated different aspects of it, with the premise that not only it is desirable for a Wikipedia language edition to contain LGBT+ articles, but also that these are framed from this perspective; in other words, that they contain explicit references to the topic.

Firstly, our approach is good to select many articles that relate to LGBT+, but sometimes also catches articles in which this relation is not perceivable. On the one hand, our manual assessment showed more false positives than false negatives for the classified LGBT+ articles in each Wikipedia language edition (average 7.2% false positives and 1.3% false negatives), which means that it is likely there is a similar percentage of false positives in this global list of unique articles and that, in reality, there could be fewer LGBT+ articles than we have collected. Adding more features to the classifier, especially ones that consider the text and the occurrence of certain terms (e.g., "gay," "lesbian," etc.), would possibly improve the accuracy of the classifier. On the other hand, the selection of the global list of unique LGBT+ articles was limited to the 87 Wikipedia language editions with the "LGBT" Wikipedia category. This means that there potentially exist unique LGBT+ articles in the rest of the Wikipedia language editions, even though they do not contain an LGBT category. This, however, makes it unlikely that these articles are numerous.

Secondly, we must acknowledge that the selection of LGBT+ articles contains a wide variety of topics and subtopics, some of which relate directly to LGBT+, while others may refer to it in a very tangential way. Once we have collected the LGBT+ articles, it would be interesting to cluster the different types of LGBT+ articles according to their relevance to the topic. This could be tackled using some of the different types of features than the ones we have used, but also, we would benefit from using the entire text of the article to identify the weight of LGBT in it, generally, and very especially in the first paragraph of the article.

Thirdly, the study of content gaps like LGBT+, gender, or geography benefit greatly those editors who are unaware of their existence, but also those who are working on reducing them and now have indicators that can help them redefine their priorities or reason their actions according to robust data. In this sense, we believe that in an open environment like Wikimedia, the results of a study should be transparent and communicated to the stakeholders as early as they are available and aim at providing real-time tools to continue observing them. In the future, we expect we can create more interactive tools like the dashboard we have shown, maybe also focusing on following the evolution of the LGBT+ content gap.[8]

Recommendations for the Wikimedia LGBT+ community

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

In this last subsection, based on the research results from this study and our knowledge of the Wikimedia Movement, we want to suggest three recommendations for the Wikimedia LGBT+ community.

1. There should be a Wikiproject, campaign, or event dedicated to completing or extending each type of data point we took advantage of for the selection of articles.

   Firstly, creating the "LGBT" category is something that only needs to be done once, even though in some communities, this idea may meet some resistance. For this reason, a global campaign and preparing local support in advance could be helpful. Secondly, adding information on the sexual orientation and partner/spouse properties in Wikidata is important as it gives clarity. The Wikidata "Wikiproject LGBT" coordinates this work and should encourage Wikimedia Movement chapters to collaborate. Thirdly, introducing links to "LGBT" in sections of geographical articles to explain the situation in terms of rights or in any general topic to include the LGBT perspective is something that is already tackled and suggested in previous research (Wexelbaum, 2019). Fourthly and lastly, the creation of articles with the term "LGBT" in the title is crucial for two reasons. It gives the reader a valuable summary of how LGBT relates to another topic. It encourages the creation of more related articles since articles containing the term “LGBT” in the title usually list so many relevant subtopics within them.

2. LGBT+ articles are visible in the lists of Featured Articles, but the Main Page is also a relevant space where they should be displayed.

   The articles that appear on the Main Page receive additional page views (Thij et al., 2019), basically because the Main Page is among the most visited pages on Wikipedia in every language edition. For this reason, there are campaigns led by gender gap groups of editors (e.g., in Spanish Wikipedia, there is "Mujeres en Portada") (“Wikiproyecto: Mujeres en Portada”, 2021) to fight for more parity in the number of biographies. Which articles appear on the Main Page is variable on the language community, and some decide that manually, with the help of algorithms that use specific features or combinations of both. In the dashboard "Home Page Gender Visibility" from the Wikipedia Diversity Observatory, there is a graph showing the number of men and women who appear on the Main Page of the 308 Wikipedia language editions on a daily basis. The LGBT+ community could push for the LGBT+ featured articles to appear on the Main Page more often so that the topics that provide important information get more attention than one or two appearances due to the international day of the Gay Pride (Wikipedia contributors, 2021i) or the death of a renowned LGBT celebrity.

3. The creation of LGBT+ articles that involve local content should become an invitation to geographical Affiliates (Wikimedia Chapters) and GLAM.

   The creation of more local LGBT+ content emerged as a future priority of the results, given that the share of local content in each language edition's LGBT+ articles is rather low, and that local content is constantly growing in each Wikipedia language (Miquel-Ribé and Laniado, 2018). In this sense, we must recognize that while some local LGBT+ information may be available, other more specific information may require access to specific databases and the collaboration of GLAM partners. The affiliate Wikimedia LGBT+ includes in its mission (1) to encourage creating partnerships with LGBT+ cultural

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

organizations, (2) to promote Wikimedia projects, and (3) to expand and create content of LGBT+ interest.

While this is perfectly aligned with local content creation, it may be difficult for a single affiliate to plan and execute strategies for every Wikipedia language edition, given that the scope is possibly too wide. For this reason, we think that growing local LGBT+ content should be among the goals of the Wikimedia chapters that are spread geographically around the globe. These chapters are in a better position to establish collaborations with LGBT+ and GLAM institutions that are an essential piece in the creation, storage, and dissemination of knowledge of interest to the LGBT+ community. The role of Wikimedia LGBT+ can be that of a supporter, providing guidance, designing strategies, and at the same time monitoring content growth with the content gaps tools and dashboards that might be derived from the results of this research.

Endnotes

[1] Wexelbaum (2019) referred to this as "queering" straight content.
[2] We can only do this selection of the LGBT+ articles in the 94 Wikipedia language editions which contain the "LGBT" category. Instead, the final selection of existing LGBT+ articles is computed for every language edition.
[3] The ground truth selection of articles is composed of articles that we assume are undoubtedly related to LGBT+.
[4] scikit provides open-source Python-based tools for predictive data analysis, available at https://www.scikit-learn.org.
[5] In this document we sometimes refer to it as ML-classifier or ML-algorithms interchangeably.
[6] In the description of the property P91 Sexual Orientation warns that "the sexual orientation of the person — use IF AND ONLY IF they have stated it themselves, unambiguously, or it has been widely agreed upon by historians after their death."
https://www.wikidata.org/wiki/Property:P91
[7] Creation date, number of Bytes, number of discussions, number of editors, number of edits, number of inlinks, number of inlinks from CCC, number of interwiki links, number of outlinks, number of outlinks to CCC, number of pageviews, number of references, number of Wikidata Properties, and LGBT indicator.
[8] A preliminary version including visualizations to depict share of LGBT+ content (and of LGBT+ biographies) in October 2020 can be seen at the Wikipedia Diversity Observatory webpage:
https://wdo.wmcloud.org/lgbt+_gap/

Acknowledgements

Andreas Kaltenbrunner acknowledges support from Intesa Sanpaolo Innovation Center. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Additional explanations of Wikimedia terms

In this section, we provide a short glossary with some Wikimedia-related terms and concepts we used throughout the study.

Edit-a-thon: An edit-a-thon (sometimes written editathon) is an event where editors of online communities such as Wikipedia improve a specific topic. The events typically include basic editing training for new editors and may be combined with a more general social meetup.

GLAM: The GLAM-Wiki initiative ("galleries, libraries, archives, and museums" with Wikipedia; also including botanic gardens and zoos) helps cultural institutions share their resources with the world through collaborative projects with experienced Wikipedia editors.

Humaniki: Humaniki is a project that extracts and visualizes data about gender, date of birth, place of birth about humans in all Wikimedia projects, typically Wikipedia biography articles.

Qitem: Qitems are Wikidata Items are also unique. Each item should represent a clearly identifiable concept or object. There is a Wikidata Qitem for every Wikipedia article, and when two or more languages have an article in common, this relates to a single Qitem.

Wikidata: "Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects, including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others" (source: Wikidata as a Linked Data Platform. https://kspurgin.github.io/presentation-20190306-wikidata-wilson-learning-forum/)

Wikiproject: "A Wikiproject is a group of contributors who want to work together as a team to improve Wikipedia. These groups often focus on a specific topic area (for example, Wikiproject Mathematics or Wikiproject India), a specific part of the encyclopedia (for example, Wikiproject Disambiguation), or a specific kind of task (for example, checking newly created pages)" (source: https://en.wikipedia.org/wiki/Wikipedia:WikiProject).

Wikimedia Affiliates: Wikimedia Foundation Board of Trustees recognizes models of affiliation within the Wikimedia movement – chapters, thematic organizations, and user groups. "Wikimedia Movement affiliates exist to further the goals of Wikimedia. Depending on their affiliation model, they do so by engaging in a wide range of activities" (from Wikipedia:WikiProject - Wikipedia. (source: https://meta.wikimedia.org/wiki/Wikimedia_movement_affiliates/Frequently_asked_questions).

Wikimedia Foundation: The Wikimedia Foundation (WMF) is an American non-profit and charitable organization that supports and participates in the Wikimedia movement, owning the internet domain names of its projects and hosting its websites.

Wikimedia LGBT+: Wikimedia LGBT+ is a Wikimedia user group that promotes the development of content on Wikimedia projects which are of interest to LGBT+ communities. The Wikimedia LGBT+ User Group was approved by the Affiliations Committee in September 2014.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Wikipedia Diversity Observatory: The Wikipedia Diversity Observatory (WDO) is a space to study diversity in Wikipedia's content and communities, identify and discuss needs and gaps, and propose and develop solutions to bridge them.

Appendix B. Additional explanations of some methodological terms

In this section, we provide short explanations of some terms and concepts related to the methodology we used throughout the study.

Ground-truth: Ground truth is a term used in various fields to refer to information that is known to be real or true, provided by definitions, user annotations, or measurements. In machine learning, it is the ideal expected result.

F1-score: The F1-score or F-measure is a measure of a test's accuracy. It is the harmonic mean of the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

Louvain Method: The Louvain method for community detection is a method to extract communities from large networks created by Blondel et al. from the University of Louvain (the source of this method's name).

Machine Learning Classifier: Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data," in order to make predictions or decisions without being explicitly programmed to do so.

Positive training set: a positive training set is a data set of positive examples used during the learning process and is used to fit the parameters (e.g., weights) of, for example, a classifier. In a binary classifier, there is usually a positive and a negative training set.

References

Ayoub, P. M., & Brzezińska, O. (2016). Caught in a web? The Internet and deterritorialization of LGBT activism. In D. Paternotte & M. Tremblay (Eds.), *The Ashgate research companion to lesbian and gay activism* (pp. 241-258). Routledge.

Blackwell, L., Hardy, J., Ammari, T., Veinot, T., Lampe, C., & Schoenebeck, S. (2016, May). LGBT parents and social media: Advocacy, privacy, and disclosure during shifting social movements. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 610-622). https://doi.org/10.1145/2858036.2858342

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008-12. https://doi.org/10.1088/1742-5468/2008/10/P10008

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Cocciolo, A. (2017). Community archives in the digital era: A case from the LGBT community. *Preservation, Digital Technology & Culture, 45*(4), 157-165. https://doi.org/10.1515/pdtc-2016-0018

Cooban, G. (2017). Should archivists edit Wikipedia, and if so how? *Archives and Records, 38*(2), 257-272. https://doi.org/10.1080/23257962.2017.1338561

Cooper, M., & Dzara, K. (2010). The Facebook revolution: LGBT identity and activism. In C. Pullen & M. Cooper (Eds.), *LGBT identity and online new media* (pp. 114-126). Routledge. https://doi.org/10.4324/9780203855430-16

Doyle, K. (2018). Minding the gaps: Engaging academic libraries to address content and user imbalances on Wikipedia. In M. Proffitt (Ed.), *Leveraging Wikipedia: Connecting communities of knowledge*, (pp. 55-69). ALA Editions.

Dyer, C. (2014). Notes on noise contrastive estimation and negative sampling. arXiv preprint https://arxiv.org/abs/1410.8251

Galloway, E., & DellaCorte, C. (2014). Increasing the discoverability of digital collections using Wikipedia: The Pitt experience. *Pennsylvania Libraries: Research & Practice, 2*(1), 84-96. https://doi.org/10.5195/palrap.2014.60

Gates, G. J. (2011). How many people are lesbian, gay, bisexual and transgender? *UCLA: The Williams Institute*. https://williamsinstitute.law.ucla.edu/publications/how-many-people-lgbt/

Gates, G. J. (2017). LGBT data collection amid social and demographic shifts of the US LGBT community. *American Journal of Public Health, 107*(8), 1220-1222. https://doi.org/10.2105/AJPH.2017.303927

Gay pride. (2021i, October 21). In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Gay_pride&oldid=1051007070

Grants: Conference/Kawayashu/Queering Wikipedia. (2021, July 31). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Grants:Conference/Kawayashu/Queering_Wikipedia&oldid=21813591.

Hawkins, B., & Watson, R. J. (2017). LGBT cyberspaces: A need for a holistic investigation. *Children's Geographies, 15*(1), 122-128. https://doi.org/10.1080/14733285.2016.1216877

Herbert, V. G., Frings, A., Rehatschek, H., Richard, G., & Leithner, A. (2015). Wikipedia–challenges and new horizons in enhancing medical education. *BMC medical education, 15*(1), 1-6. https://doi.org/10.1186/s12909-015-0309-2

Humaniki | Wikimedia Diversity Dashboard Tool. (2021). *Gender by language editions in Wikimedia Projects*. Retrieved October 27, 2021, from https://humaniki.wmcloud.org/gender-by-language

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Jemielniak, D., Masukume, G., & Wilamowski, M. (2019). The most influential medical journals according to Wikipedia: Quantitative analysis. *Journal of Medical Internet Research, 21*(1), e11429. https://doi.org/10.2196/11429

Knowledge Equity Calendar/1/en. (2019, December 2). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Knowledge_Equity_Calendar/1/en&oldid=19604398

Knowledge Equity Calendar/15/en. (2019, December 20). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Knowledge_Equity_Calendar/15/en&oldid=19652650

Konieczny, P., & Klein, M. (2018). Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator. *New Media & Society, 20*(12), 4608-4633. https://doi.org/10.1177%2F1461444818779080

LGBT+ articles dashboard. (2021b). https://wdo.wmcloud.org/lgbt+_articles

List of articles every Wikipedia should have. (2021, September 14). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=List_of_articles_every_Wikipedia_should_have&oldid=22017101

marcmiquel/WDO. (2021). *WDO/lgbt_content_selection.py at wcdo*. Retrieved October 27, 2021, from https://github.com/marcmiquel/WDO/blob/wcdo/src_data/lgbt_content_selection.py

Mehra, B., & Srinivasan, R. (2007). The library-community convergence framework for community action: Libraries as catalysts of social change. *Libri, 57*(3), 123-139. https://doi.org/10.1515/LIBR.2007.123

Miquel-Ribé, M., & Laniado, D. (2018). Wikipedia culture gap: Quantifying content imbalances across 40 language editions. *Frontiers in Physics, 6*, 54. https://doi.org/10.3389/fphy.2018.00054

Miquel-Ribé, M., & Laniado, D. (2020). The Wikipedia Diversity Observatory: A project to identify and bridge content gaps in Wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration* (pp. 1-4). https://doi.org/10.1145/3412569.3412866

Moodie, C. (2016, January 11). *David Bowie's wild love life: How the boy kept swinging with a string of men and women*. Mirror.co.uk. https://www.mirror.co.uk/3am/celebrity-news/david-bowies-wild-love-life-7161395

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology, 65*(12), 2381-2403. https://doi.org/10.1002/asi.23162

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830. https://arxiv.org/abs/1201.0490

Person Task Force. (2021, November 23). In *Wikipedia*. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_LGBT_studies/Task_forces/Person

Phetteplace, E. (2015). Accidental technologist: How can libraries improve Wikipedia? *Reference & User Services Quarterly, 55*(2), 109. http://dx.doi.org/10.5860/rusq.55n2.109

Portal: LGBT. (2021d, October 22). In *Wikipedia, The Free Encyclopedi*a. https://en.wikipedia.org/w/index.php?title=Portal:LGBT&oldid=1051300274

Pullen, C., & Cooper, M. (Eds.). (2010). *LGBT identity and online new media*. Routledge.

Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. (2020). A Taxonomy of knowledge gaps for Wikimedia projects (Second Draft). arXiv preprint. https://arxiv.org/abs/2008.12314

Soriano, C. R. R. (2014). Constructing collectivity in diversity: Online political mobilization of a national LGBT political party. *Media, Culture & Society*, *36*(1), 20-36. https://doi.org/10.1177/0163443713507812

Stewart, B., & Kendrick, K. D. (2019). "Hard to find": Information barriers among LGBT college students. *Aslib Journal of Information Management, 71*(5), 601-617. https://doi.org/10.1108/AJIM-02-2019-0040

Strategy/Wikimedia movement/2017/Direction. (2021). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Strategy/Wikimedia_movement/2017/Direction&oldid=21540194

Szajewski, M. (2013). Using Wikipedia to enhance the visibility of digitized archival assets. *D-Lib Magazine, 19*(3). https://doi.org/10.1045/march2013-szajewski

Thij, M., Kaltenbrunner, A., Laniado, D., & Volkovich, Y. (2019). Collective attention patterns under controlled conditions. *Online Social Networks and Media, 13*, 100047. https://doi.org/10.1016/j.osnem.2019.07.003

Warncke-Wang, M., Ranjan, V., Terveen, L., & Hecht, B. (2015, April). Misalignment between supply and demand of quality content in peer production communities. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 9, No. 1). https://ojs.aaai.org/index.php/ICWSM/article/view/14631

Wexelbaum, R. (2019). Coming out of the closet: Librarian advocacy to advance LGBTQ+ Wikipedia engagement. In *LGBTQ+ Librarianship in the 21st Century: Emerging Directions of Advocacy and Community Engagement in Diverse Information*

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

Environments (Advances in Librarianship, Vol. 45), Emerald Publishing Limited, Bingley, pp. 115-139. https://doi.org/10.1108/S0065-283020190000045011

Wexelbaum, R., Herzog, K., & Rasberry, L. (2015). Queering Wikipedia. In R. Wexelbaum (Ed.), *Queers online: LGBT digital practices in libraries, archives, and museums* (pp. 61–80). Litwin Books.

Wiki99. (2021, July 6). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Wiki99&oldid=21702112

Wikimedia Foundation. (2021). Wikimedia Downloads. https://dumps.wikimedia.org/

Wikimedia LGBT+/Portal. (2021, October 24). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Wikimedia_LGBT%2B/Portal&oldid=22233292

Wikimedia movement. (2021). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Wikimedia_movement&oldid=22035811

Wikipedia contributors. (2021a). Wikipedia: GLAM. In Wikipedia, The Free Encyclopedia. Retrieved 09:28, October 27, 2021, https://en.wikipedia.org/w/index.php?title=Wikipedia:GLAM&oldid=1026460753

Wikipedia Diversity Observatory. (2021, August 2). In *Meta, discussion about Wikimedia projects*. https://meta.wikimedia.org/w/index.php?title=Wikipedia_Diversity_Observatory&oldid=21827818

Wikipedia: Categorization/Ethnicity, gender, religion, and sexuality. (2021e, October 13). In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Categorization/Ethnicity,_gender,_religion_and_sexuality&oldid=1049779586

Wikipedia: Featured articles. (2021f, October 27). In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Featured_articles&oldid=1052035729

Wikipedia: Featured articles. (2021g, October 27).In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Featured_articles&oldid=1052035729

Wikipedia: Vital articles. (2021h, October 27).In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Vital_articles&oldid=1052080224

Wikipedia: Wikipedians (2021b, September 28). In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedians&oldid=1047027878

Wikipedia: WikiProject LGBT studies. (2021c, September 21). In *Wikipedia, The Free Encyclopedia*.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270

https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_LGBT_studies&oldid=1044119150

Wikiproyecto: Mujeres en Portada. (2021, February 7). In *Wikipedia, La Enciclopedia Libre*.https://es.wikipedia.org/w/index.php?title=Wikiproyecto:Mujeres_en_Portada&oldid=133029869

Wulczyn, E., West, R., Zia, L., & Leskovec, J. (2016, April). Growing Wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 975-985).

Marc Miquel-Ribé (mmiquel-ctr@wikimedia.org) is a university professor and PhD researcher based in Barcelona (Catalonia). He teaches user experience at the Universitat Pompeu Fabra (UPF) - Tecnocampus and does research on content diversity and editor engagement in online communities. He has been a member of Amical Wikimedia (Catalan Wikipedia) since 2011. Additionally, he's been one of the lead writers of the Wikimedia Strategy 2030 Plan and helped shape the narrative to prioritize equity and inclusion in future movement projects. He is currently working in the Wikimedia Foundation research team on the project Knowledge Gaps Index and in partnership with the Eurecat Foundation in a project named Wikipedia Community Health Metrics.

Andreas Kaltenbrunner (kaltenbrunner@gmail.com) is Senior Research Scientist at the ISI Foundation in Turin. He holds a PhD in Computer Science and Digital Communication obtained in 2008 from the UPF in Barcelona. Afterwards, he has worked at the technology centre Barcelona Media where he co-founded the Social Media research line and led it from May 2013 onwards. Between June 2015 and August 2017, he was Scientific Director of the Digital Humanities Research Unit at the technology centre Eurecat. In September 2017 he joined NTENT as Director of Data Analytics, until joining ISI Foundation in October 2020. Andreas is also teaching a master course on data based social analytics at Universitat Pompeu Fabra (UPF) in Barcelona and is involved in research activities centred on computational social science, social media and social network analysis. He has co-authored more than 70 publications in these areas.

Jeffrey M. Keefer (jk904@nyu.edu) is an open learning and non-profit capacity building consultant, educational and institutional researcher, professor, and Wikimedian. He has worked in higher education and organizational learning for nearly two decades, and helps people navigate their learning needs and take informed action.

ISSN 2574-3430, jps.library.utoronto.ca/index.php/ijidi/index
DOI: 10.33137/ijidi.v5i4.37270